MBS TECHNICAL REPORT MBS 15-06

Decomposing Models of Bounded Rationality

Daniel Jessie*and Ryan Kendall†

June 16, 2015

Abstract

This paper illustrates a general disconnection between many models of bounded rationality and human decision making. A new mathematical approach allows for any game to be decomposed into unique components. The "strategic" component of a game contains the necessary and sufficient information to determine the prediction for a broad class of models focused on bounded rationality. Among others, this class of models includes the most commonly used specifications for Quantal Response (QRE), Noisy Introspection (NI), level-k, and Cognitive Hierarchy (CH). These bounded rationality models are shown to exhibit a mathematical invariance to changes in a game's non-strategic components, and this paper's primary hypothesis is that humans do not exhibit this invariance. Using a laboratory experiment consisting of simple 2×2 games, we find that human subjects systematically respond a game's "behavioral" component, which is ignored by the QRE, NI, level-k, and CH models. We show that previous results and puzzles related to these models are special cases of our general finding. In addition, our approach can predict the settings in which contemporaneous models of bounded rationality will generate good (and poor) fits of human behavior *before* the data is collected, making it a valuable tool for future research.

*International Institute for Applied Systems Analysis (IIASA). Laxenburg, Austria. Email: jessie@iiasa.ac.at

†Postdoctoral Research Associate. Department of Economics. University of Southern California. Los Angeles Behavioral Economics Laboratory (LABEL). Email: rakendal@usc.edu

1 Introduction

What will people choose in the following simple games?

Game A

	L	R
T	19, 22	4, 19
B	14, 3	3, 1

Game B

	L	R
T	6, 5	21, 2
B	1, 24	20, 22

The Nash Equilibrium (Nash, 1950 and 1951) of both games is Top-Left. However, should we expect that the choice made by the row player in game A will be the same choice made by the row player in game B? What about the choice made by the column player in both games. Should we also expect these choices to be the same? Probably not. One may sense that these games are sufficiently different and, therefore, one anticipates different choices across games. One may further suspect that the row player in game A is more likely to choose Top than does the row player in game B, and the column player in game A is more likely to choose Left than does the column player in game B. With this intuition, we should expect to observe the Top-Left cell (Nash equilibrium) more often in game A than in game B.[1]

Which components of a game motivate humans to make different choices? This natural inquiry drives much of the current work in game theory and experimental economics. Field and laboratory experiments have produced a bounty of evidence suggesting that humans rarely exhibit the perfectly-discerning self-interested behavior suggested by the Nash equilibrium. In an effort to understand the divergence between observed behavior and perfectly rational models, *bounded rationality* has emerged as a key concept. Certain models of this type predict the outcome of a game as the equilibrium of error-prone decision makers,[2] while other models predict the outcome based on a measure of agents' "strategic sophistication".[3] Some models generate predictions using a combination of these features.[4] Arguably, the success of these models is shown in their ability to accurately fit data generated by human subjects typically in a laboratory setting. Indeed, the literature is growing with results and discussions emphasizing which model fits what data better.[5]

From the arsenal of bounded rationality models, which one best accounts for the obvious

[1] This straightforward intuition is shown to be accurate in our experiment.

[2] Quantal Response Equilibrium, McKelvey & Palfrey (1995); Heterogeneous Quantal Response Equilibrium, McKelvey, Palfrey, & Weber (2000); Asymmetric Logit Equilibrium, Weizsäcker (2003).

[3] Level-k, Stahl & Wilson (1994) and Nagel (1995); Cognitive Hierarchy, Camerer, Ho, & Chong (2004).

[4] Noisy Introspection, Goeree & Holt (2004); Truncated Quantal Response Equilibrium, Rogers, Palfrey, & Camerer (2009).

[5] Crawford, Costa-Gomes, & Iriberri (2013) survey this growth with a particular focus on nonequilibrium models of strategic thinking. Section 2 of this paper more thoroughly discusses the previous research modeling bounded rationality.

difference between games A and B? The surprising fact illustrated in this paper is that none of the commonly used bounded rationality models predict different human behavior between games A and B.[6] This suggests that the current debate over which model fits what data more accurately is overlooking a general disconnection between the components that influence *all* of these bounded rationality models and the components that influence human choices.

The innovative aspect of this paper's experimental design is that it utilizes a new approach for analyzing a game as the sum of its components. Using this approach, described in detail in Jessie & Saari (2013), we can provide a characterization of which components are relevant to our current models of bounded rationality, along with which components are irrelevant. If a component is irrelevant to a model, than any change to this component will not alter the fit of that model. All models are invariant to some components in a game, and certain invariances could be desirable. For example, if we do not expect humans to behave differently in games where monetary values are measured in dollars or in dimes, then a model that is invariant to this scaling component of a game is desirable. However, if humans systematically respond to a component of the game that a model is invariant to, this disconnection will determine the model's fit to be *predictably* different from the observed human behavior.

As section 3 more precisely describes, we use the approach developed by Jessie & Saari in order to uniquely decompose any 2×2 game into two main components; what we refer to as *strategic* and *behavioral*.[7] This decomposition is useful because a broad class of bounded rationality models generate their prediction using solely the strategic component and are, therefore, invariant to changes in the behavioral component.[8] By holding constant the strategic component and changing only the behavioral component, we are able to design games that are mathematically equivalent from the viewpoint of many bounded rationality models (such as games A and B).The implication of these models is that human behavior will be constant across games with the same strategic component, and our primary hypothesis is that it will not be constant.

Section 4 describes a laboratory experiment that we use in order to test this implication. We find that human subjects do not exhibit the predicted invariance to the behavioral component.[9] In particular, subjects in the experiment are shown to systematically respond to the behavioral component. These findings, presented in section 5, suggest that human subjects

[6]This includes models with logistically specified errors and/or assuming that level-0 players choose randomly. This result also holds for a general class of model specifications, but these two are overwhelmingly favored in applications of these models.

[7]A scaling or *kernel* component is the third and final component of any game.

[8]Section 3 identifies the bounded rationality models that solely rely on the strategic component.

[9]The use of "prediction" here does not refer to a model estimation based on observed data–which is a fit–but refers instead to an ex-ante qualitative feature of the model. In this case, a feature resulting from a particular mathematical invariance.

respond to games in a qualitatively different way than many contemporary model predict.

The importance of this result rests in its applications to future research and previous findings. The findings of this paper suggest a particular class of games in which we can expect our current models to provide accurate predictions, which implies that future projects using such models should only apply them to this restricted class of games. Furthermore, we are able to unify puzzles found earlier in the literature as unique cases of our general result. These main avenues in which we contribute to the literature are described in section 6. Section 7 addresses limitations of this paper and section 8 concludes.

2 Literature Review

The typical definition of perfectly rational actions in a strategic setting is given by the Nash equilibrium. In an attempt to characterize non-Nash human behavior, two main traits of the Nash equilibrium have been relaxed: best-responding and belief-choice consistency. Here, we briefly discuss models that relax best-responding, models that relax equilibrium from belief-choice consistency, and models that relax both traits.

Relaxing perfectly rational best-responding behavior is typically done by assuming that agents choose an action according to a probabilistic choice mechanism that determines them to be "better"-responders. There are many interpretations as to why humans would behave according to this mechanism. One common interpretation of this behavior is that, rather than perfectly observing the payoff from a strategy (either π_1 or π_2), an agent perceives the payoff as the true payoff disturbed by a shock, ε. The salience of this shock can be controlled by a magnitude parameter, μ. With this stochastic process, an agent compares the disturbed payoffs rather than the true payoffs. For example, the agent chooses strategy 1 if

$$\pi_1 + \mu\varepsilon_1 > \pi_2 + \mu\varepsilon_2. \tag{1}$$

There are many interpretations as to why it is plausible to model agents with this noisy process (risk aversion, attention, strategic sophistication, and so on). Regardless of the interpretation of ε, strategy 1 is selected if

$$\frac{(\pi_1 - \pi_2)}{\mu} > \varepsilon_2 - \varepsilon_1. \tag{2}$$

The probability that this inequality holds can be expressed as $F\left(\frac{(\pi_1-\pi_2)}{\mu}\right)$, where F$(\cdot)$ is the distribution function of the difference in the shocks. When the shocks are assumed to be identically and independently drawn from a type-I extreme-value distribution, an agent's

probabilistic choice is modeled by the familiar logit equation.[10] This process generates a probabilistic best response function for an agent in which strategies that yield higher payoffs are more likely (but not necessarily) selected than strategies that yield lower payoffs. Using this error-specification, the probability that Agent i selects Strategy s from the possible strategy set S is given by the following equation.

$$p_s^i = \frac{e^{\lambda E(\tilde{\pi}_s^i(p^{-i}))}}{\sum_{s'}^{S} e^{\lambda E(\tilde{\pi}_{s'}^i(p^{-i}))}} \tag{3}$$

While λ (or $\frac{1}{\mu}$) can be interpreted in many ways, this parameter simply controls the salience of the shocks that agents experience in the decision process. Equation 3 displays Agent i's logit quantal response function where the fixed point of every agent's logit quantal response function is the logit Quantal Response Equilibrium (McKelvey & Palfrey, 1995, QRE hereafter). While the logit QRE imposes a homogeneous λ for all agents, other models allow for a separate λ value for each agent (Heterogeneous Quantal Response Equilibrium, McKelvey, Palfrey, & Weber, 2000, HQRE hereafter; Asymmetric Logit Equilibrium, Weizsäcker, 2003, ALE hereafter). For homogeneous or heterogeneous λ values, the intuition is the same: agents noisily respond to the possible payoffs of a game, but still achieve an equilibrium solution based on their correct belief about their competitor's level of noise (λ). Therefore, these bounded rationality models relax best-responding but preserve the assumption of equilibrium choices.

There exists another strand of literature that retains agents who best-respond but relaxes the assumption of equilibrium choices. The most common models of this type are level-k (Stahl & Wilson, 1994; Nagel, 1995) and Cognitive Hierarchy (Camerer, Ho, & Chong, 2004, CH hereafter). In these models, agents have different levels of sophistication and each agent's choice is based on their belief about their competitors' sophistication level. In level-k, an agent with sophistication level $n > 0$ perfectly best-responds as if she were playing against a level $n-1$ opponent. In Cognitive Hierarchy, this level-n agent best-responds as if she were playing against a distribution of agents with lesser sophistication levels (typically Poisson distributed). The prediction for both level-k and Cognitive Hierarchy are dependent on the analyst's assumption about the behavior of the level-0 agent. Similar to the probabilistic models, the typical assumption for setting the level-0 agent's behavior is to assume perfect randomness. That is, level-0 will select any available strategy with equal probability. Furthermore, agents are allowed to have beliefs about their competitors that are inconsistent with their competitors' actual choice, thus relaxing Nash's equilibrium

[10] The long literature using this type of model starts builds on the classical work by Luce (1959) and McFadden (1973).

constraint.

A third strand of literature relaxes both best-responding and equilibrium. The models in this literature have better-responding probabilistic agents who are not constrained to have a belief-choice consistency about their opponents. The two well-known models in this literature are Noisy Introspection (Goeree & Holt, 2004, NI hereafter) and the Truncated Quantal Response Equilibrium (Rogers, Palfrey, & Camerer, 2009, TQRE hereafter).

3 A General Invariance

While the models described in section 2 use different approaches to model bounded rationality, this section illustrates that all of these models exhibit the same invariance. To do so, we rely on the mathematical representation of $k_1 \times k_2 \times \cdots \times k_n$ games developed in Jessie & Saari (2013), and this is the unique representation that captures this invariance. Section 3.1 briefly describe this approach applied to 2×2 games to provide the reader with intuition as to why the decomposition works. Section 3.2 demonstrates that the specified models mentioned in section 2 calculate their fit based solely on the strategic component of a game and are, therefore, invariant to changes in the remaining non-strategic components.

3.1 Decomposing Games

Any $k_1 \times k_2 \times \cdots \times k_n$ game $\mathcal{G}$ can be viewed as consisting of three components: a strategic component, which determines the Nash best-response; a behavioral component, which influences a variety of features, such as Pareto efficiency; and a kernel component, which adds a constant to each of an agent's payoffs. We apply this approach to a 2×2 game where agent "Row" selects either Top or Bottom and agent "Column" selects either Left or Right. For example, if we are interested in analyzing how agents choose strategies in game A, we can decompose this original game into the following components.

Game A	L	R		*strategic*			*behavioral*			*kernel*	
T	19, 22	4, 19	=	2.5, 1.5	0.5, −1.5	+	6.5, 9.25	−6.5, 9.25	+	10, 11.25	10, 11.25
B	14, 3	3, 1		−2.5, 1	−0.5, −1		6.5, −9.25	−6.5, −9.25		10, 11.25	10, 11.25

The term on the far right of the equation is the kernel component. The values in each cell of this component can be found by taking the average of an agent's payoff over each possible outcome. So for Row, the kernel value in each cell is $k^R = (19+4+14+3)/4 = 10$. Column's kernel value is $k^C = 11.25$. In any 2×2 game, each agent has one kernel value that is represented in all possible cells of the kernel component.

The middle term is the behavioral component. The behavioral values in each cell of this component can be found by computing the difference between the overall game average (kernel) and the average payoff for an agent if her opponent were to chose one action. For example, consider Row who has an overall game average is 10. Row's average payoff if Column chooses Left is $(19 + 4)/2 = 16.5$ which is different from the overall game average by 6.5. Similarly, Row's average payoff if Column chooses Right is $(4 + 3)/2 = 3.5$ which is different from the overall game average by -6.5. For 2×2 games, it is generally true that the behavioral values for any agent are the same magnitude but have different signs. Because of this, we say that each agent has one behavioral value which is represented as either the positive or negative version. In game A, Row's behavioral component is $b^R = 6.5$ which is represented as either 6.5 or -6.5 in each cell.

The payoff that an agent receives from the kernel component is the same regardless of the outcome of a game. In addition, whether an agent receives the positive or negative behavioral value is independent of her choice; whether Row receives 6.5 or -6.5 from the behavioral component depends only on Column's choice. Therefore, the information contained in the behavioral and kernel components have no effect on a strategic agent's choice. All the information expressing how Row's choice affects her own payoff is contained in the remaining, strategic, component. The values in this component capture the payoff differences between the agent's payoff in a specific cell and that agent's average payoff over all possible strategies if her opponent were to chose one action. For example, if Column chooses Left, then Row has a choice between receiving 19 from Top and 14 from Bottom. The average payoff in this case is $(19 + 14/2) = 16.5$, which is already captured by the sum of the behavioral and kernel components. The only remaining information is the difference from the average, 2.5 for Top and -2.5 for Bottom. In this way, each agent in a 2×2 game has one strategic value (one positive and one negative) for each of her opponent's possible actions. This means that agents will have two strategic values.[11] Importantly, cells within the strategic component that have positive strategic values for both agents represent outcomes in the original game where neither agent can increase their personal payoff by deviating, or pure Nash equilibria of the game.

This approach of decomposing games reveals that the strategic component of game B is the exact same as the strategic component of game A. This means that, if people choose differently in game A as they do in game B (as we suspect they will), this difference will be solely driven by differences in non-strategic components (behavioral and kernel).

The general decomposition of a 2×2 game is shown below. As is shown by Jessie &

[11] In fact, for 2×2 games, the entire Nash equilibrium structure of any game is determined by a single statistic from each player, and can be represented by points in a unit square (Jessie & Saari, 2013).

Game B	L	R		*strategic*			*behavioral*			*kernel*	
T	6, 5	21, 2	=	2.5, 1.5	0.5, −1.5	+	−8.5, −9.75	8.5, −9.75	+	12, 13.25	12, 13.25
B	1, 24	20, 22		−2.5, 1	−0.5, −1		−8.5, 9.75	8.5, 9.75		12, 13.25	12, 13.25

Saari (2013), this is the unique way to decompose a game so that the strategic information is completely separate from the non-strategic (behavioral and kernel) information.

Game $\mathcal{G}$	L	R		*strategic*			*behavioral*			*kernel*	
T	$\pi_1^R,\ \pi_1^C$	$\pi_2^R,\ \pi_2^C$	=	$s_1^R,\ s_1^C$	$s_2^R,\ -s_1^C$	+	$b^R,\ b^C$	$-b^R,\ b^C$	+	$k^R,\ k^C$	$k^R,\ k^C$
B	$\pi_3^R,\ \pi_3^C$	$\pi_4^R,\ \pi_4^C$		$-s_1^R,\ s_2^C$	$-s_2^R,\ -s_2^C$		$b^R,\ -b^C$	$-b^R,\ -b^C$		$k^R,\ k^C$	$k^R,\ k^C$

Figure 1. General decomposition of a 2×2 game. The superscripts denote the agent and subscripts are an index.

3.2 Invariance to non-strategic components

The strategic component is the only component that captures the relationship between an agent's choice and that agent's personal payoff. In a game, agents best-respond by comparing the expected payoffs for each strategy given their belief about their opponent and selecting the strategy that yields the highest expected payoff. Consider Row's comparison over Top or Bottom against Column who she believes is choosing Left with probability q. Equations 4 represent this standard comparison using the general notation from the game payoffs (Game $\mathcal{G}$) depicted in Figure 1.

$$EV(Top) = q\pi_1^1 + (1-q)\pi_2^1$$

$$EV(Bottom) = q\pi_3^1 + (1-q)\pi_4^1 \tag{4}$$

These equations are used to provide a risk-neutral Row's best-response correspondence defined for any belief about the choice made by Column. Row will choose Top if $EV(Top|q) > EV(Bottom|q)$, Bottom if $EV(Top|q) < EV(Bottom|q)$, and will be indifferent over any mixture of Top and Bottom if $EV(Top|q) = EV(Bottom|q)$. In some games, Column will have an available mixing strategy of Left and Right that would make Row indifferent between choosing Top or Bottom: q^*. In these games, this mixture, along with a similarly derived p^*, represent the mixed Nash Equilibrium of the game.

$$q^* = \frac{\pi_4^R - \pi_2^R}{\pi_1^R - \pi_2^R - \pi_3^R + \pi_4^R} \tag{5}$$

$$p^* = \frac{\pi_4^C - \pi_3^C}{\pi_1^C - \pi_2^C - \pi_3^C + \pi_4^C} \tag{6}$$

In order to show that an agent's best response correspondence is entirely captured by the strategic component, we can perform this standard analysis using the decomposed components of a game. Rather than using the game payoffs, insert the decomposed values into Equation 4 to get the following expected payoffs.

$$EV(Top) = q[s_1^R + b^1 + k^1] + (1-q)[s_2^R - b^1 + k^1] = qs_1^R + (1-q)s_2^R + [q(b^1 + k^1) + (1-q)(k^1 - b^1)]$$

$$EV(Bottom) = q[-s_1^R + b^1 + k^1] + (1-q)[-s_2^R - b^1 + k^1] = -qs_1^R - (1-q)s_2^R + [q(b^1 + k^1) + (1-q)(k^1 - b^1)] \tag{7}$$

The equations have been rearranged to illustrate that the expected payoff of Top and the expected payoff of Bottom have the same bracketed factor at the end of each expression. Importantly, the bracketed factor includes all of the behavioral and kernel values and none of the strategic values. Therefore, any comparison of these expected payoffs will not be influenced by the behavioral or kernel components; the expected values differ only in the strategic terms. Furthermore, note that the mixed Nash equilibrium is found to only depend on strategic values.

$$q^* = \frac{-s_2^R}{s_1^R - s_2^R} \tag{8}$$

$$p^* = \frac{-s_2^C}{s_1^C - s_2^C} \tag{9}$$

Because best-response correspondences are invariant to changes in non-strategic components, all Nash equilibria of a game can be found using only the strategic component. Perhaps surprisingly, this result extends to models that allow agents to have imperfect beliefs about the strategy chosen by their opponents such as level-k and CH. In both models, an agent with a sophistication-level greater than 0 will best-respond to their belief about their opponent. For example, Row with sophistication-level greater than 0 will develop a belief about the probability that Column will choose L, $\hat{q}$. Row will then compare the expected payoffs in Equation 7 using this belief and, as shown above, this comparison only relies on the strategic component. However, there exists an additional nuance in these models that is absent in the Nash framework. In level-k or CH, the analyst's specification of the level-0 agent will influence the $\hat{q}$ used by higher level agents to best-respond and, because of this, a model's invariance to non-strategic components depends this specification. This means that specifying the level-0 agents in such a way that they respond to the kernel or behavioral component will

produce models that are not invariant to changes in the non-strategic components. Examples of this type of specification are level-0 agents that always select the strategy that contains the outcome yielding the highest possible payoff (no matter how unlikely they are to receive it), or level-0 agents that always select the strategy with the highest amount of even numbers. These types of specifications are typically not employed, and are not being tested in this paper, as they are not invariant to changes in non-strategic components. However, there are many level-0 specifications that do produce models with such an invariance. In general, any specification that assumes level-0 agents only respond to the strategic component of a game or something payoff independent will be invariant to changes in non-strategic components. The most prominent example of this type is to specify level-0 agents to select one action from their strategy set with uniform probability. Importantly, level-k or CH models that assume this type of "random" level-0 agent will be invariant to changes in non-strategic components, and are subject to our experimental analysis.[12]

Nash, level-k, and CH are best-responding models that solely rely on the strategic component. Because the strategic component also captures the cardinal difference in expected payoffs between strategies, this result extends further to include models that assume error-prone agents making choices according to better-response functions.[13] While this result holds for models with an array of different error-specifications, we focus our analysis on the most commonly used specification where errors are the product of a logistic process.[14] Following from Equation 3, we can express Row's better-response function using the decomposed values in a general 2×2 game.[15]

$$p_T^R = \frac{e^{\lambda E(\tilde{\pi}_T^R)}}{e^{\lambda E(\tilde{\pi}_T^R)} + e^{\lambda E(\tilde{\pi}_B^R)}}$$

$$p_T^R = \frac{e^{\lambda\left(qs_1^R + (1-q)s_2^R + q(b^R+k^R) + (1-q)(k^R-b^R)\right)}}{e^{\lambda\left(qs_1^R + (1-q)s_2^R + q(b^R+k^R) + (1-q)(k^R-b^R)\right)} + e^{\lambda\left(-qs_1^R - (1-q)s_2^R + q(b^R+k^R) + (1-q)(k^R-b^R)\right)}}$$

$$p_T^R = \left(\frac{e^{\lambda\left(q(b^R+k^R) + (1-q)(k^R-b^R)\right)}}{e^{\lambda\left(q(b^R+k^R) + (1-q)(k^R-b^R)\right)}}\right) \frac{e^{\lambda\left(qs_1^R + (1-q)s_2^R\right)}}{e^{\lambda\left(qs_1^R + (1-q)s_2^R\right)} + e^{\lambda\left(-qs_1^R - (1-q)s_2^R\right)}}$$

[12] In addition to this commonly used specification, there are other level-0 specifications that are also invariant to changes in the behavioral component. For instance, this is true if level-0 agents always select their first strategy (Top for Row and Left for Column) or if level-0 agents follow a heuristic pattern such as "choose Top on even rounds and choose Bottom on odd rounds" or "random round 1 and mimic whatever my opponent chooses thereafter".

[13] This idea was originally pioneered by the highly influential QRE by McKelvey & Palfrey (1995).

[14] More generally, better-responding models will be invariant to non-strategic components if the errors are specified so that each agent's possible strategies have the same expectation over the error term. For example, in a 2×2 game, this would impose that $E(\varepsilon_1|Top) = E(\varepsilon_1|Bottom)$ for Row. An error specification violating this relatively weak assumption would assume that agents are more likely to choose certain strategies as a direct result of the structure of the error terms. We do not focus on these types of unusually specified models.

[15] This equation is analogous to the logit quantal response function for Row.

$$p_T^R = \frac{e^{\lambda\left(qs_1^R+(1-q)s_2^R\right)}}{e^{\lambda\left(qs_1^R+(1-q)s_2^R\right)} + e^{\lambda\left(-qs_1^R-(1-q)s_2^R\right)}} \tag{10}$$

Column has a similar better-response function that depends on p, s_1^C, and s_2^C. As with the best-response correspondences, these better-response functions do not contain the behavioral or kernel component, which means their fit will solely rely on the strategic component.

In the logit-QRE, agents are assumed to have the same level of responsiveness to error-shocks (homogeneous λ value) and each agent's belief about her opponent is assumed to align with that opponent's actual strategy (equilibrium play). The resulting prediction is the fixed point of two agents making choices according to equations of the form of Equation 10. The logit-QRE is clearly invariant to changes in non-strategic components and this mathematical invariance extends to other better-responding models that relax a common λ parameter and/or equilibrium play. The logit-HQRE and ALE models assume agents in the same game can have different levels responsiveness to error-shocks (heterogeneous λ values). Since the strategic component defines the response function for all possible levels of responsiveness, a model allowing for heterogeneous λ values will still be invariant to changes in non-strategic components. Therefore, the logit-HQRE and ALE are also invariant to changes in non-strategic components.

The logit-NI model allows for an agent to select a strategy based on their belief about their opponent's strategy which may not align with their opponent's actual strategy (non-equilibrium play). In this model, unlike the logit-QRE, logit-HQRE, or ALE, an agent is not constrained to choose the strategy at the fixed point of the better-response functions. However, agents in the NI model are still constrained to select a strategy along the response function, which is solely influenced by the strategic component. Therefore, the logistically specified NI model is invariant to changes in the non-strategic components. The iterative manner in which the NI model relaxes belief-choice consistency of the QRE is similar to the manner in which level-k relaxes belief-choice consistency of the Nash Equilibrium.

Finally, the logit-TQRE allows for heterogeneous λ values and non-equilibrium play. Neither of these assumptions, alone, produce a model that considers a part of the game other than the strategic component. Therefore, the logit TQRE is also invariant to changes in the non-strategic components. The downward looking distributional belief process of the TQRE model relaxes belief-choice consistency of the QRE in a similar manner to the way in which the CH relaxes belief-choice consistency of the Nash Equilibrium. Indeed, Rogers, Palfrey, & Camerer (2009) show that the CH is a special case of the TQRE.

The main point of this section is to illustrate that the mathematics underlying models that use these better-response functions will classify games as equivalent if they have the same strategic components. Intuitively, this is the same type of mathematical equivalence observed when a calculator treats the expressions 1 divided by 2, 3 divided by 6, or 26 divided by 52 as all mathematically equivalent to 0.5. The calculator processes the components of 1 and 2, or 3 and 6, or 26 and 52 into the information that the calculator deems as essential–namely, the ratio between the two components. As we know, there are an infinite number of ways to combine such components in order to represent 0.5 in this manner. This intuition is reflected in the way that better-responding models classify games. There exists a continuum of games that only differ in the behavioral or kernel component but all have the same strategic component (e.g. games A and B). Better-responding models process these components into the information that they deem as essential–the strategic component. Also, there are an infinite number of games that can be created that are mathematically equivalent to each other in this way. This analogy is helpful for understanding the mechanism determining the model's fit as well as for developing the appropriate strategy for estimating the fit of the observed data (discussed in section 5).

4 Experiment Design

Section 3 demonstrates that the Nash, Level-k, CH, logit-QRE, ALE, logit-HQRE, logit-NI, and logit-TQRE concepts have the same mathematical invariance to non-strategic components of a game. The fit of these specified models is found using only the strategic component of a game. This is important for our purpose because fixing the strategic component while changing the behavioral and kernel component will not alter the outcome of the models. The natural question is whether human subjects make the same choices in games that have the same strategic component. In other words, *do humans exhibit the invariance predicted by many models of bounded rationality*?

4.1 Strategically equivalent sets

In order to address this question, we designed an experiment testing whether or not humans will respond to differences in the behavioral component of games. To do so, we construct five-game "sets" where the non-strategic information is the only variation between games in the same set. Figure 2 depicts one of the sets used in the experiment. Because every 2×2 game in a set will have the exact same strategic component, we label these sets of games as "strategically equivalent sets". Each game in a strategically equivalent set will also have a very

similar kernel component. The sole purpose of the slight variation in the kernel component is to avoid negative payoffs, fractional payoffs, and subject-recognition of similar games. This flexibility allows us avoid confounding issues such as loss-aversion, confusion, or learning. Given that the kernel component adds the same term to a subject's payoff in all possible outcomes, the small changes that we employ to this component are likely to be innocuous. We do not expect the small variations in the kernel component to have a meaningful impact on the choices made in our experiment.[16] Furthermore, beyond the strategically equivalent sets, the kernel component is held relatively constant across all 30 games. The only meaningful difference between games within the same strategically equivalent set is their behavioral component. These sets allow us to identify human responsiveness to the behavioral component by observing a difference in the choices made in 2×2 games within the same strategically equivalent set.

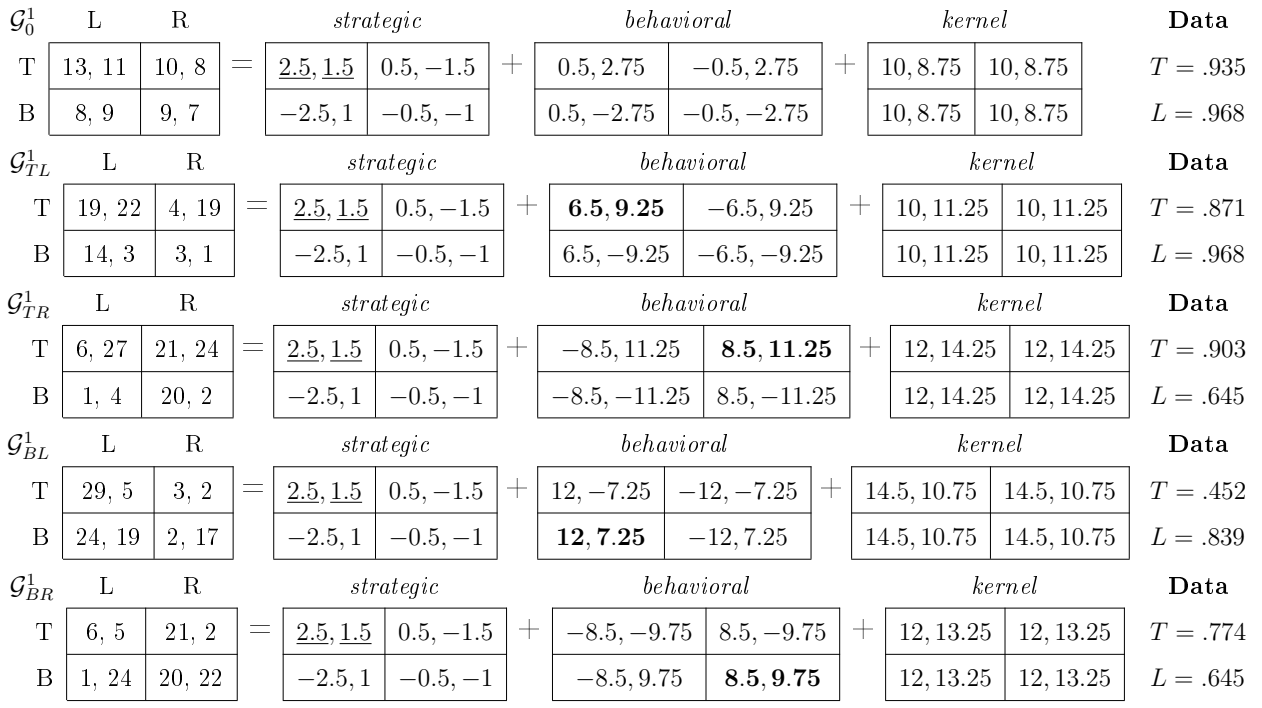

Figure 2. Strategically equivalent set #1. This figure shows the original game displayed to the subjects along with each game's decomposition and the proportion that a strategy was observed in the experiment. Each game is labeled with a superscript denoting the strategic equivalent set in which it belongs and a subscript denoting the bias of the behavioral component.

[16] However, it is interesting to consider the reaction of human choices to very large difference in the kernel component. Very large differences in the kernel component will not be captured by many models but it is plausible that humans could respond to such differences.

4.2 Hypotheses

This paper's primary goal is to illustrate *any* difference between human choices in games that belong to the same strategically equivalent set. In this manner, our primary hypothesis is defined below.

> *Primary Hypothesis.* Subjects will choose different actions in games that belong in the same strategic equivalent set.

In addition, our design allows for testing a more particular hypothesis focused on the predictable way in which subjects respond to the behavioral component. Not surprisingly, results from game theory and economics experiments have illustrated that subjects will be predictably influenced by the strategic component of each game. In particular, they will be likely to select an action that has a positive strategic value for themselves resulting in a higher personal payoff. When both subjects act in this manner, the outcome of the game will align with one cell in the strategic component where both values are positive. Such cells represent a game's pure Nash Equilibrium and, as such, 2×2 games can either have zero, one, or two cells where both subjects have positive strategic values.[17] In our experiment, we posit that subjects will be influenced by the behavioral component in a similar way. Since a subject's choice in the behavioral component does not influence their personal payoff, we hypothesize that subjects will be more likely to select an action that has the positive behavioral value for the other subject. If both subjects act in this manner, then the outcome of the game will align with the cell in the behavioral component where both values are positive. Unlike the strategic component, the behavioral component always has only one cell in which both subjects have behavioral values greater than or equal to zero. All of the behavioral values used in our experiment are non-zero which means that every game will have one unique cell in which both subjects have positive behavioral values. We anticipate that a higher proportion of the outcomes will be in the cell of the original game that corresponds with this "biased" cell in the behavioral component where both values are positive.

> *Secondary Hypothesis.* In the biased games, subjects will be more likely to select an action that has the positive behavioral value for the other subject.

For example, consider games A and B (which correspond to games $\mathcal{G}^1_{TL}$ and $\mathcal{G}^1_{BR}$ in strategically equivalent set #1). The cell in which both behavioral values are positive in games A and B are the Top-Left and Bottom-Right cell, respectively. Because of this, our secondary hypothesis expects Top and Left to be chosen more often in game A along with Bottom and Right to be chosen more often in game B. In our experiment, four games within a strategically equivalent set have a relatively large biased behavioral component that has

[17] This corresponds to games with one mixed Nash Equilibrium, games with one pure Nash Equilibrium, and games that have two pure Nash Equilibria and one mixed Nash Equilibrium, respectively.

both positive values in either the Top-Left, Top-Right, Bottom-Left, or Bottom-Right. These biased games will be respectively labeled $\mathcal{G}_{TL}, \mathcal{G}_{TR}, \mathcal{G}_{BL}, \mathcal{G}_{BR}$. The remaining game, $\mathcal{G}_0$ has a relatively unbiased behavioral component that is comprised of smaller values to serve as a baseline for which to test the responsiveness for the other four games in that set.

4.3 Types of games

Strategically equivalent set #1 is comprised of five games where both subjects have dominant strategies that align with the single pure Nash Equilibrium: Top for Row and L for Column. Each game's pure Nash Equilibrium is represented in the strategic component by the cell with the underlined payoffs.[18] The behavioral component plays the primary role in differentiating each game. Each game's bias is represented in the behavioral component with bold payoffs. In order to test our hypotheses, the bulk of our analysis will compare the choices observed in the set's biased games ($\mathcal{G}_{TL}, \mathcal{G}_{TR}, \mathcal{G}_{BL}, \mathcal{G}_{BR}$) with the choices observed in that set's unbiased game ($\mathcal{G}_0$). This analysis is presented in the following section.

In order to test the robustness of human-responsiveness to a game's behavioral component, our experiment includes strategically equivalent sets with games that have one pure Nash equilibrium where both subjects have dominant strategies (Set #1 and #2), games with multiple Nash equilibria (Set #3 and #4), games with no pure-strategy Nash equilibrium (Set #5), and games with one pure Nash equilibrium where one subject has a dominant strategy and the other subject does not (Set #6). While the Nash structure of each strategically equivalent set is different, each set is made up of one unbiased game and four biased games and the order in which the biased games are labeled is also the same as in strategically equivalent set #1. The experiment consists of six different strategically equivalent sets leading to a total of 30 2×2 games.[19] Table 1 describes the Nash structure of each of the six strategically equivalent sets and Appendix A contains the games used in the experiment along with each game's decomposition and data from the experiment.

4.4 Experiment Procedures

Our data consist of two experimental sessions totaling 62 subjects. These subjects were recruited using online software from the subject pool for the Experimental Social Science

[18] Because each game has the same strategic component, all Nash equilibria and the prediction of the previously-mentioned bounded rationality concepts will be the same for every game in a set.

[19] The experiment also included a simple 2×2 stag hunt game either played by the subject at the very beginning or very end of the experiment. This was done in order to test a hypothesis outside the scope of this paper. As with all the other games in the experiment, subjects were not informed about the outcome of this game until the experiment was over. It is very unlikely that this game had any effect on behavior.

Set	N.E. count	Type of games
1	$(p = 1, q = 1)$	Dominant strategies for both subjects
2	$(p = 1, q = 0)$	Dominant strategies for both subjects
3	$(p = 1, q = 1), (p = 0, q = 0), (p = \frac{3}{5}, q = \frac{1}{3})$	Battle of Sexes (coordination)
4	$(p = 1, q = 0), (p = 0, q = 1), (p = \frac{1}{3}, q = \frac{3}{8})$	Battle of Sexes (anti-coordination)
5	$(p = \frac{4}{7}, q = \frac{1}{3})$	Matching Pennies
6	$(p = 1, q = 0)$	Dominant strategy for one subject

Table 1. The different Nash structures used in the experiment.

Laboratory at the University of California, Irvine. The subjects went through an instruction phase along with practice questions to ensure comprehension. Subjects were told they would be paid according to the average dollar amount earned in 5 randomly chosen games. Each subject made a one-time simultaneous choice in each of the 30 randomly displayed 2×2 games using the z-tree software (Fischbacher, 2007). Subjects in the experiment were only presented with the composed game and were never informed about how games could be decomposed. No cells or values were underlined, bolded, or emphasized in any manner. Feedback about their opponent's identity, their opponent's choice, or the outcome of each game was not provided for the subjects until all 30 games were completed. This was done in order to discourage subjects from developing different strategies as the experiment progressed due to learning or coordination efforts. In addition, since our experiment did not provide any information about other subjects during the experiment, we believe that the behavior of one subject could not have a large influence on the behavior of all of the subjects in an entire session.

Furthermore, while subjects made choices according to the 2×2 matrices shown in Appendix A, every subject viewed the games as if they were playing from the perspective of Row choosing between Top or Bottom. For example, when two subjects are matched to play game A, one subject is randomly determined to view the game as if it were game A whereas the other subject views the game as if it were game A'.

Game A

	L	R
T	19, 22	4, 19
B	14, 3	3, 1

Game A'

	L	R
T	22, 19	3, 14
B	19, 4	1, 3

In this manner, all subjects were making 30 choices over Top or Bottom. Of course the viewpoint of the subject has no mathematical effect on an agent's choice and we are not testing any effect stemming from this approach. This design feature was employed in order to make the subject's choice as straightforward as possible.

5 Results

Each 2×2 game has two different subjects each faced with two different strategy choices (The row subject chooses over Top or Bottom and the column subject chooses over Left or Right). The data points that we analyze are the aggregated strategy choices within each 2×2 game. These data points are represented as the observed aggregated proportion of subjects who chose Top (for the row subjects) or the observed aggregated proportion of subjects who chose Left (for the column subjects). Each game's observed strategy is aggregated from the 31 subjects who made a choice in that game as either the row subject or the column subject. While this is a relatively small sample size, we achieve enough statistical power to fully support our primary hypothesis - that humans respond to differences in the behavioral component of games and are, therefore, not invariant to the behavioral component. Furthermore, we find that responsiveness is robust to all of the Nash-structures we tested in this experiment. In addition, our data partially support our secondary hypothesis - that humans follow the behavioral component's cell where both values are positive. This human responsiveness is especially systematic in games with one pure Nash Equilibrium. These results are formally presented in the remainder of this section.[20]

In order to analyze human-responsiveness to the behavioral component, we focus on the differences between the observed choices in games that belong to the same strategically equivalent set.[21] Within each set we test for statistical differences between the observed choices in the game with the unbiased behavioral component with the observed choices in each of the four games with a biased behavioral component. Naturally, we only compare the choices made by the row subjects with other choices made by the row subjects. For example, in strategically equivalent set #1 we test for a significant difference between the aggregated row subject's choice between $\mathcal{G}_0^1$ and $\mathcal{G}_{TL}^1$, $\mathcal{G}_0^1$ and $\mathcal{G}_{TR}^1$, $\mathcal{G}_0^1$ and $\mathcal{G}_{BL}^1$, and $\mathcal{G}_0^1$ and $\mathcal{G}_{BR}^1$. Similar tests are performed for the column subject, which determines 8 tests for each strategically equivalent set. With six strategic equivalent sets, in total we perform 48 two-sample two-sided t-tests testing for differences in the observed proportions. The p-values of these tests are displayed in Table 2.

[20] Our experimental results, in and of themselves, are not surprising because it aligns with one's intuition that subjects will behave differently in games within the same strategically equivalent set (consider games A and B from the introduction). The crux of this paper is that many models determine their predicted fit without considering this obvious difference between games. Because of this, we expect that a replication of our study with more subjects would simply find the same qualitative results supporting our two hypotheses.

[21] Comparing games across sets would conflate human-responsiveness to the behavioral with responsiveness to the strategic component.

			$\mathcal{G}^1_{TL}$		$\mathcal{G}^1_{TR}$		$\mathcal{G}^1_{BL}$		$\mathcal{G}^1_{BR}$	
			T	L	T	L	T	L	T	L
		obs.	.871	.968	.903	.645	.452	.839	.774	.645
$\mathcal{G}^1_0$	T	.935	.3946	-	.6443	-	.0000	-	.0722	-
	L	.968	-	1.00	-	.0013	-	.0854	-	.0013

			$\mathcal{G}^2_{TL}$		$\mathcal{G}^2_{TR}$		$\mathcal{G}^2_{BL}$		$\mathcal{G}^2_{BR}$	
			T	L	T	L	T	L	T	L
		obs.	.968	.387	.871	.129	.645	.258	.581	.161
$\mathcal{G}^2_0$	T	.968	1.00	-	.1604	-	.0013	-	.0003	-
	L	.097	-	.0077	-	.6907	-	.0971	-	.4522

			$\mathcal{G}^3_{TL}$		$\mathcal{G}^3_{TR}$		$\mathcal{G}^3_{BL}$		$\mathcal{G}^3_{BR}$	
			T	L	T	L	T	L	T	L
		obs.	.613	.645	.742	.258	.355	.742	.290	.419
$\mathcal{G}^3_0$	T	.645	.7942	-	.4075	-	.0224	-	.0051	-
	L	.806	-	.1555	-	.0000	-	.5469	-	.0018

			$\mathcal{G}^4_{TL}$		$\mathcal{G}^4_{TR}$		$\mathcal{G}^4_{BL}$		$\mathcal{G}^4_{BR}$	
			T	L	T	L	T	L	T	L
		obs.	.484	.613	.581	.258	.452	.645	.194	.226
$\mathcal{G}^4_0$	T	.548	.6141	-	.7933	-	.4497	-	.0039	-
	L	.290	-	.0106	-	.7776	-	.0051	-	.5647

			$\mathcal{G}^5_{TL}$		$\mathcal{G}^5_{TR}$		$\mathcal{G}^5_{BL}$		$\mathcal{G}^5_{BR}$	
			T	L	T	L	T	L	T	L
		obs.	.613	.742	.742	.645	.161	.452	.419	.290
$\mathcal{G}^5_0$	T	.903	.0077	-	.0971	-	.0000	-	.0001	-
	L	.387	-	.0048	-	.0421	-	.6041	-	.4196

			$\mathcal{G}^6_{TL}$		$\mathcal{G}^6_{TR}$		$\mathcal{G}^6_{BL}$		$\mathcal{G}^6_{BR}$	
			T	L	T	L	T	L	T	L
		obs.	.871	.774	.871	.258	.613	.613	.645	.323
$\mathcal{G}^6_0$	T	.935	.3946	-	.3946	-	.0024	-	.0051	-
	L	.419	-	.0044	-	.1804	-	.1264	-	.4340

Table 2. P-values from two-sided t-tests testing for statistical difference between the strategies observed in $\mathcal{G}_0$ and the strategies observed in games $\mathcal{G}_{TL}, \mathcal{G}_{TR}, \mathcal{G}_{BL}, \mathcal{G}_{BR}$.

As was shown in section 3, many models of bounded rationality are invariant to the difference between games within the same strategically equivalent set. This implies that none of the 48 comparisons between choices should be different. However, 21 out of 48 strategy

comparisons were different at a significance level of .05. Furthermore, these differences are not relegated to certain Nash structures, but rather exist in every Nash structure tested in this paper. Every strategic equivalent set has (at least) three out of the four games in which (at least) one subject chooses a different strategy than the unbiased game at the .05 level. This result supports our primary hypothesis that humans are responsive to a general aspect of games that is ignored by a large class of bounded rationality models.

To illustrate the fundamental difference between model fit and human behavior, we estimated the predicted fit of the logit QRE using our experimental data. Because the model's parameter value, λ, is implicitly tied to the Nash structure of the game, it would be a misuse of the concept to jointly estimate one parameter using the choice data collected in different strategic equivalent sets. Therefore, we do not fit the data by estimating the same λ parameter over all 30 games used in the experiment.[22] Different studies estimate λ over different Nash-structures in an effort to achieve parsimonious models and to avoid over-fitting. Indeed, many positive results pertaining to the accuracy of the logit QRE in experimental settings have been found by estimating λ over different Nash-structured games (Goeree & Holt 2005, Levine & Palfrey 2007, and Selten & Chmura 2008). This approach can be attractive in certain settings because it prohibits the model from artificially fitting the data as the product of a highly specified model. In our analysis, however, we allow for this level of over-fitting by estimating an individual λ for each strategically equivalent set. Estimating the same λ for all five games within a strategically equivalent set is the appropriate approach because the logit QRE is invariant to changes in the behavioral or kernel component. So from the point of view of the logit QRE, all five games within a strategically equivalent set are mathematically equivalent.[23] This estimation strategy yields a 6 parameter model for estimating the choices made in 30 games. A maximum likelihood approach is used to estimate λ values for each strategically equivalent set.[24] These parameters determine the logit QRE fit for each game in Figure 2 which separately compares the fit and observations for each subject. The y-axis is either the logit QRE's predicted fit or the observed probability that an action is chosen (Top by the row subject and Left by the column subject). The x-axis represents the 30 different games. Games that belong to the same strategically equivalent set are represented as data points connected with a line.[25]

[22]However, estimating one λ for all 30 games will produce the same qualitative illustration as in Figure 2. This estimation strategy would produce one flat-line fit for all 30 games, which would also illustrate the logit QRE's unresponsiveness to the behavioral component.

[23]The logit QRE's sole reliance on the strategic component is mathematically shown in Equation 10 and intuitively illustrated with the previously mentioned analogy relating these models to a calculator.

[24]Strategic equivalent sets $1, 2, 3, 4, 5$, and 6 have an estimated λ of $.379, .352, -.038, .017, -.005$, and $.352$, respectively.

[25]Refer to Table 1 for the Nash-structure of each strategically equivalent set and to Figure 2 along with Appendix A for a decomposition of all 30 games.

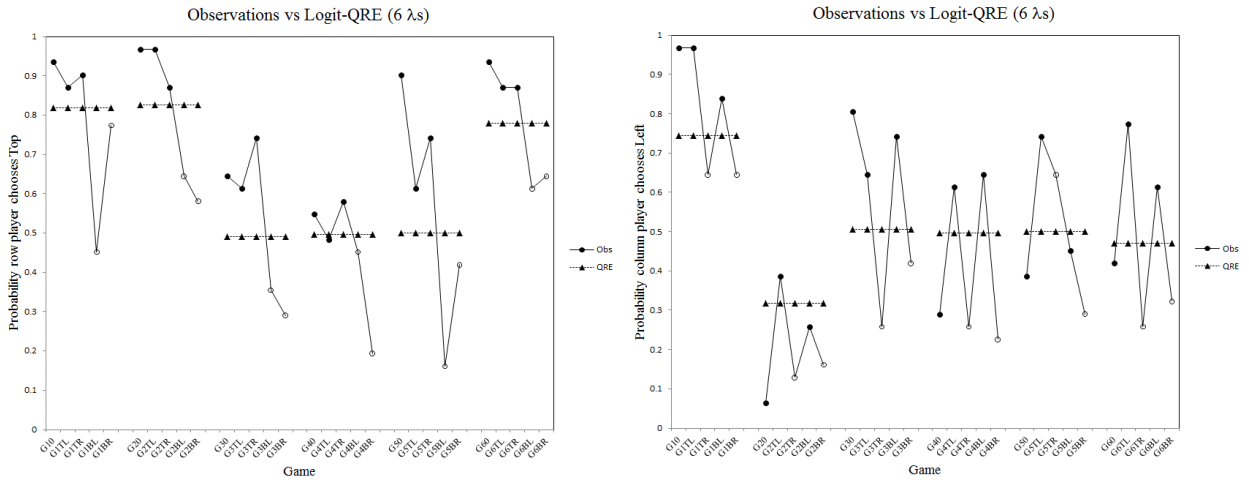

Figure 2. This figure shows aggregated observed strategy choice in each of the 30 games along with the estimated fit using the logit QRE. This figure separately illustrates this comparison for the row subject (left) and the column subject (right). Empty data circles represent games where the behavioral component was biased in the Bottom or Right action. The secondary hypothesis would predict that these empty data circles would be lower on the y-axis.

As Figure 2 illustrates, the fit offered by the logit QRE fails to capture any of the the observed differences in human behavior within strategically equivalent sets. The six-parameter logit QRE generates horizontally "flat-line" fits within each strategically equivalent set. This unresponsive fit is the expected result based on the logit QRE's invariance to the behavioral component of games (as described in section 3). Furthermore, all of the bounded rationality models shown to have the same invariance in section 3 of this paper will offer the same type of flat-line fit between strategically equivalent sets (albeit possibly different flat-lines). The general implication is that human subjects do not exhibit the same invariance as do the mathematical models, resulting in a large difference between the predicted fit (absolutely no change in behavior within strategically equivalent sets) and the observation (large differences within strategically equivalent sets).

Our secondary hypothesis posits that subjects will be more likely to select the action where the behavioral component for the other subject is positive. This implies that all 48 comparisons would be statistically significant in the predicted direction. In general, this is not true as 21 of the comparisons are significant at the .05 level. However, we do find support for this hypothesis in two important ways.

First, this hypothesis is perfectly supported when restricting the analysis to settings in which (i) a subject has a dominant strategy and (ii) the behavioral component is biased towards a different cell than the Nash Equilibrium. Both subjects in strategically equivalent sets #1 and #2 have dominant strategies as well as the row subject in strategically equivalent set #6. Consider strategically equivalent set #1 where the row and column subjects have

dominant strategies to choose Top and Left, respectively. While the behavioral component in game $\mathcal{G}^1_{TL}$ is biased, it is biased toward the same strategies as each subject's dominant strategy (and the subsequent pure Nash Equilibrium at Top-Left). The bias introduced in game $\mathcal{G}^1_{TL}$ aligns with the strategic component, and only serves to reinforce each subject's dominant strategy, which they were already overwhelmingly favoring. This serves as one explanation for why we should not expect that the strategies observed in these two games will be statistically different from each other. However, the other games within these sets are biased in such a manner that one or both of the subjects should be expected to change their choice in a predicted manner. In game $\mathcal{G}^1_{TR}$, we observe evidence that the subjects were systematically responding to the behavioral component because the column subject selected Right more often than in game $\mathcal{G}^1_0$ (p-value of .0013) and the row subject's strategy was unchanged. We observe the same effect in game $\mathcal{G}^1_{BL}$ where the row subject's strategy favored Bottom more often than in game $\mathcal{G}^1_0$ (p-value of .0000) and the column subject's strategy remained relatively unchanged. In game $\mathcal{G}^1_{BR}$, the row subject favors Bottom (p-value of .0722) and the column subject favors Right (p-value of .0013) more often than they do in $\mathcal{G}^1_0$. This trend is also observed for any subject who was presented with a dominant strategy and conflicting strategic and behavioral components (both subjects in strategically equivalent set #2 and the row subject in strategically equivalent set #6).

Second, our data also support a weaker version of our secondary hypothesis that is captured in all of the strategic equivalent sets. Rather than testing for differences between the biased games and the unbiased game within a set, a more relaxed comparison would be to test for differences between the biased games that conflict with each other. Are decisions different in games with a Top-biased behavioral component versus games with a Bottom-biased behavioral component? What about Left-biased versus Right-biased? This test can be illustrated by comparing the filled and empty data circles in Figure 2. The filled data circles represent games where the behavioral component is biased toward the Top or Left action, whereas the empty data circles represent games where the behavioral component is biased toward the Bottom or Right action. With this illustration, our Secondary Hypothesis would predict that the observations with filled data circles would be higher on the y-axis for both subjects, which is what is observed in Figure 2. Within each strategically equivalent set, the row subject selected the Top action significantly more often in the average of the two top Top-biased games than they did versus the average of the two Bottom-biased games. This was true for all six strategically equivalent sets for the row subject (Table 3). Similarly, the column subject is more likely to select the Left strategy in Left-biased games in five out of the six strategically equivalent sets (Table 4). These results can be visualized in Figure 2 as tests for significant differences between the average of the two filled in circles and the

average of the two empty circles within each strategically equivalent set.

Set #1		Set #2		Set #3	
$\frac{\mathcal{G}^1_{TL}+\mathcal{G}^1_{TR}}{2}$	$\frac{\mathcal{G}^1_{BL}+\mathcal{G}^1_{BR}}{2}$	$\frac{\mathcal{G}^2_{TL}+\mathcal{G}^2_{TR}}{2}$	$\frac{\mathcal{G}^2_{BL}+\mathcal{G}^2_{BR}}{2}$	$\frac{\mathcal{G}^3_{TL}+\mathcal{G}^3_{TR}}{2}$	$\frac{\mathcal{G}^3_{BL}+\mathcal{G}^3_{BR}}{2}$
0.887	0.613	0.919	0.613	0.677	0.322
Diff (p-value)= .0004		Diff (p-value)= .0000		Diff (p-value)= .0000	
Set #4		Set #5		Set #6	
$\frac{\mathcal{G}^4_{TL}+\mathcal{G}^4_{TR}}{2}$	$\frac{\mathcal{G}^4_{BL}+\mathcal{G}^4_{BR}}{2}$	$\frac{\mathcal{G}^5_{TL}+\mathcal{G}^5_{TR}}{2}$	$\frac{\mathcal{G}^5_{BL}+\mathcal{G}^5_{BR}}{2}$	$\frac{\mathcal{G}^6_{TL}+\mathcal{G}^6_{TR}}{2}$	$\frac{\mathcal{G}^6_{BL}+\mathcal{G}^6_{BR}}{2}$
0.532	0.322	0.678	0.290	0.871	0.629
Diff (p-value)= .0183		Diff (p-value)= .0000		Diff (p-value)= .0019	

Table 3. Row subject's behavior in Top-biased games and Bottom-biased games within each strategic equivalent set. P-values represent two-sided t-tests testing for statistical difference.

Set #1		Set #2		Set #3	
$\frac{\mathcal{G}^1_{TL}+\mathcal{G}^1_{BL}}{2}$	$\frac{\mathcal{G}^1_{TR}+\mathcal{G}^1_{BR}}{2}$	$\frac{\mathcal{G}^2_{TL}+\mathcal{G}^2_{BL}}{2}$	$\frac{\mathcal{G}^2_{TR}+\mathcal{G}^2_{BR}}{2}$	$\frac{\mathcal{G}^3_{TL}+\mathcal{G}^3_{BL}}{2}$	$\frac{\mathcal{G}^3_{TR}+\mathcal{G}^3_{BR}}{2}$
0.903	0.645	0.323	0.145	0.694	0.339
Diff (p-value)= .0006		Diff (p-value)= .0196		Diff (p-value)= .0001	
Set #4		Set #5		Set #6	
$\frac{\mathcal{G}^4_{TL}+\mathcal{G}^4_{BL}}{2}$	$\frac{\mathcal{G}^4_{TR}+\mathcal{G}^4_{BR}}{2}$	$\frac{\mathcal{G}^5_{TL}+\mathcal{G}^5_{BL}}{2}$	$\frac{\mathcal{G}^5_{TR}+\mathcal{G}^5_{BR}}{2}$	$\frac{\mathcal{G}^6_{TL}+\mathcal{G}^6_{BL}}{2}$	$\frac{\mathcal{G}^6_{TR}+\mathcal{G}^6_{BR}}{2}$
0.629	0.242	0.597	0.468	0.694	0.290
Diff (p-value)= .0000		Diff (p-value)= .150		Diff (p-value)= .0000	

Table 4. Column subject's behavior in Left-biased games and Right-biased games within each strategic equivalent set. P-values represent two-sided t-tests testing for statistical difference.

6 Application

The main message from our experiment is that humans systematically respond to a component of games that is ignored by a large class of bounded rationality models. This result is particularly powerful because we can predict this behavior before any data is generated or collected. This section presents two ways in which such an ex-ante approach can contribute to our understanding of human behavior. First, we provide a framework that can be applied to future research projects focused on bounded rationality or social preferences. Second, we relate our finding - that humans respond to the behavioral component of games - to previous puzzles in the literature by suggesting that the inconsistencies found in this early work represent special cases of our general result.

6.1 Future projects

The decomposition provides an ex-ante framework detailing when current models of bounded rationality are most likely to be accurate predictors of human behavior. While outside the scope of this paper, additional experiments could be used to validate our suggested predictive framework described in this subsection. In general, we should expect that these models will perform well in settings where one of three conditions is satisfied: (i) the bias of the strategic and behavioral components are completely aligned, (ii) the strategic component is large relative to the behavioral (and kernel) component(s), or (iii) the behavioral (and kernel) component(s) are held constant across games.[26]

The bias of the strategic and behavioral components are completely aligned if each component has the same (one) cell where the value for each agent is positive. Certain games in this paper serve as examples of condition (i), as the observed behavior was unchanged between biased games that completely align with the strategic component in that strategic equivalent set's unbiased game. Games $\mathcal{G}_0^1$, $\mathcal{G}_0^2$, $\mathcal{G}_0^6$ have strategic components that define the game to have a unique Nash Equilibrium in the Top-Left, Top-Right, and Top-Right cell, respectively. In our experiment, we observe no statistical difference between these three behaviorally unbiased games and the behaviorally-biased games that align with the Nash Equilibrium; $\mathcal{G}_{TL}^1$, $\mathcal{G}_{TR}^2$, $\mathcal{G}_{TR}^6$. In fact, in strategically equivalent sets with one pure Nash Equilibrium (#1, #2, and #3), these differences represent the only biased games in which we do not observe a statistical difference. Importantly, because the behavioral component for any game is necessarily biased toward one unique cell (ignoring games with 0 behavioral values), condition (i) can only be satisfied by games with one pure Nash Equilibrium. Games with multiple Nash Equilibria will have strategic components that are biased toward multiple cells, and games with no pure Nash Equilibrium will have a strategic component that is not biased toward any cell.

The games in our experiment violate condition (ii) because they all have relatively small strategic components (compared to the kernel and behavioral components). However, condition (ii) can easily be satisfied by scaling the behavioral and kernel component of any game to make the strategic component relatively larger. For example, consider new games $\tilde{A}$ and $\tilde{B}$ which are constructed by taking games A and B and dividing their behavioral and kernel components by 100.

[26] Conversely, as was shown in this paper, we should expect that these models will provide a poor fit of human behavior in settings where none of these conditions are met.

Game $\tilde{A}$	L	R
T	2.665, 1.705	0.535, -1.295
B	-2.335, 1.02	-0.465, -0.98

Game $\tilde{B}$	L	R
T	2.535, 1.535	0.705, -1.465
B	-2.465, 1.23	-0.295, -0.77

Since games A, B, $\tilde{A}$, and $\tilde{B}$ have the same strategic component, current models of bounded rationality will offer the same fit of human behavior across all four games. However, while we expect (and experimentally observe) humans to behave differently in games A and B, we expect human behavior to be relatively equivalent in games $\tilde{A}$ and $\tilde{B}$. Games $\tilde{A}$ and $\tilde{B}$ still have strategic and behavioral components that are misaligned, but the scaled down non-strategic components have a reduced impact on human behavior. Condition (ii) suggests that current models of bounded rationality will provide an accurate fit of human behavior when the strategic component dominates the non-strategic components.

Finally, we suspect that contemporaneous models of bounded rationality can accurately fit the difference in human behavior across games that hold the behavioral and kernel component constant. For instance consider games P and Q, both of which have have relatively large behavioral components that conflict with the strategic component (violating conditions (i) and (ii)).

Game P	L	R		*strategic*			*behavioral*			*kernel*	
T	40, 40	10, 16	=	$12, 12$	$-3, -12$	+	$7.5, 7.5$	$-7.5, 7.5$	+	$20.5, 20.5$	$20.5, 20.5$
B	16, 10	16, 16		$-12, -3$	$3, 3$		$7.5, 7.5$	$-7.5, -7.5$		$20.5, 20.5$	$20.5, 20.5$

Game Q	L	R		*strategic*			*behavioral*			*kernel*	
T	31, 31	1, 25	=	$3, 3$	$-12, -3$	+	$7.5, 7.5$	$-7.5, 7.5$	+	$20.5, 20.5$	$20.5, 20.5$
B	25, 1	25, 25		$-3, -12$	$12, 12$		$7.5, 7.5$	$-7.5, -7.5$		$20.5, 20.5$	$20.5, 20.5$

Because the only difference between games P and Q is the strategic component, we suspect that any observed difference between the two games is largely the result of responsiveness to this difference, which is captured by all the current models of bounded rationality. These models would predict that Top and Left would be chosen more often in game P than in game Q.[27] Indeed, this prediction aligns with the expected behavior in these two games.

This paper's decomposition represents the first complete manner in which experiments can be designed that deliberately hold constant the prediction of many personal-payoff bounded rationality concepts, but this framework also extends to research focused on modeling social preferences. Future experimental projects concerned with analyzing human responsiveness to non-strategic factors (such as many types of social preferences) can design games that take

[27] The mixed Nash Equilibria of games P and Q are $(p = \frac{4}{5}, q = \frac{4}{5})$ and $(p = \frac{1}{5}, q = \frac{1}{5})$, respectively.

advantage of this tool. Specifically, when testing for social preferences over many games, each game should have the exact same strategic component and only vary in the behavioral or kernel component. Since the structure of the strategic component was previously unknown, it is likely that many papers which focused on social preferences varied the strategic component in conflation with their desired treatment conditions. We present the first experimental analysis of this kind in Appendix B when analyzing common models of social preferences. While this analysis is outside the hypotheses of this paper, it illustrates that common models of social preferences do not provide a meaningful explanation of our data.

6.2 Previous findings

Understanding that humans will respond to the behavioral component of games can be used to add clarity to previous papers that fit one (or many) of these models of bounded rationality. In particular, inconsistencies or puzzles found in the previous literature can be unified under the general theme in this paper's results–namely, that humans respond to the behavioral component of games. In this section, we focus on two prestigious examples of this kind: Deck (2001) and Goeree & Holt (2004).

Deck (2001) analyzes the fit between many types of models and human behavior in two specific types of games: an "Exchange" game (E) and an "Investment" game (I).

Game E

	C	D
X	4, 6	4, 6
E	8, 12	0, 20

$=$ *strategic*

$-2, 0$	$2, 0$
$2, -4$	$-2, 4$

$+$ *behavioral*

$2, -5$	$-2, -5$
$2, 5$	$-2, 5$

$+$ *kernel*

$4, 11$	$4, 11$
$4, 11$	$4, 11$

Game I

	C	D
X	4, 12	4, 12
E	8, 12	0, 20

$=$ *strategic*

$-2, 0$	$2, 0$
$2, -4$	$-2, 4$

$+$ *behavioral*

$2, -2$	$-2, -2$
$2, 2$	$-2, 2$

$+$ *kernel*

$4, 14$	$4, 14$
$4, 14$	$4, 14$

Deck correctly states that the QRE model implies the same human behavior in these two games "independent of the value of λ" (p. 1550). This is clear when we see that both games have the same strategic component. One of the main results of Deck's paper is that human subjects selected action D significantly more often in game I than in game E.[28] This result, along with others, is leveraged to conclude that "the models currently discussed in the profession do not capture behavior in a broad sense." (p. 1554). Using our ex-ante decomposition approach, we can see that both games have a biased behavioral component in the C-E cell. However, the main difference between the two games is that the behavioral component in game I is smaller than the behavioral component in game E. Because game

[28] D is observed at a frequency of 0.60 and 0.29 in games I and E, respectively.

I is constructed with a smaller behavioral bias toward the C-E cell, the strategic component is relatively larger in game I than in game E. Because of this, we expect that humans will be more responsive to the strategic component in game I, which may be an additional explanation for why the weakly dominant strategy D is chosen more often in this game. Deck's result stems from his observed difference between games I and E. We expand on this idea here by illustrating a method to construct an infinite number of games that would have the same result.

Goeree & Holt's 2004 paper introduces the highly influential NI model. In this paper, they task human subjects to play the following "game of chicken" (CK) described below.

Game CK

	S	R
S	12, 12	15, 32
R	32, 15	-5, -5

$=$ *strategic*

$-10,-10$	$10,10$
$10,10$	$-10,-10$

$+$ *behavioral*

$8.5,8.5$	$-8.5,8.5$
$8.5,-8.5$	$-8.5,-8.5$

$+$ *kernel*

$13.5,13.5$	$13.5,13.5$
$13.5,13.5$	$13.5,13.5$

While the NI model fits much of their data, Goeree & Holt illustrate game CK as one example where "the introspection model predicts poorly" (p. 379). They state that "In this case, the best response functions intersect at the center of a graph... The effect of adding noise is to round off the corners, leaving S-shaped logit response functions that still intersect in the center. This symmetry causes the symmetric logit and introspection equilibra to also be at 0.5... The data, in contrast to all three predictions, reveal that 67% of choices were the safe decision [S]. This suggests that the high rate of safe choices may be due to risk aversion." (p. 379). Goeree & Holt's explanation of the Nash, QRE, NI models having a 0.5 prediction can be verified by the symmetric nature of the strategic component. The strategic component is perfectly balanced and the QRE and NI will offer the same 0.5 prediction for all levels of rationality and introspection. This prediction conflicts with what Goeree & Holt observe in their experiment. However, instead of relying on a story of risk-aversion, the behavioral component stands out as a different explanation for their observed behavior. The behavioral values are both positive in the cell where both subjects choose S. Since game CK is biased into the S-S cell, we expect subjects will choose the S action more often than is predicted by Nash, QRE, and NI (which aligns with their observations). Furthermore, using our approach, we can explicitly test whether humans are responding to the behavioral component (as we suggest) or to the level of risk associated with each action (as suggested by Goeree & Holt). An illustration of this approach is shown below in game CK'.

Game CK'

	S	R
S	21, 21	6, 23
R	23, 6	4, 4

$=$ *strategic*

$-1,-1$	$1,1$
$1,1$	$-1,-1$

$+$ *behavioral*

$8.5,8.5$	$-8.5,8.5$
$8.5,-8.5$	$-8.5,-8.5$

$+$ *kernel*

$13.5,13.5$	$13.5,13.5$
$13.5,13.5$	$13.5,13.5$

What percentage of subjects will choose S in game CK'? If risk aversion is the correct explanation of the original game CK, then we should expect very different human behavior in game CK' than was observed in game CK. Unlike in game CK, the level of risk associated with each strategy is very similar (S provides either a payoff of 21 or 6 while R provides either a payoff of 23 or 4). Because of this, if subjects are motivated to avoid risk, they will slightly prefer S to R in game CK', and the observed percentage of S choices in game CK' will be very close to (but slightly higher than) the 0.5 prediction offered by Nash, QRE, and NI. However, if subjects are motivated by the behavioral component, they will continue to illustrate a defined preference for S over R in game CK'. We suggest that subjects will, indeed, show a defined preference for S over R in game CK' as they did in the original game CK. A formal verification of this hypothesis along with a more general analysis of decomposing models of risk is further explored in Jessie & Kendall (2015).

7 Limitations

This paper is subject to two main criticisms based on assumptions made about agents being risk neutral and personal-payoff maximizers. While these are common modeling assumptions, there exist many circumstances in which humans show a preference for avoiding risk or a preference for "social" aspects such as altruism or payoff equity. As we see it, relaxing these assumptions is a fruitful venture located outside the main scope of this paper. However, both limitations are partially addressed in this section and more thoroughly addressed either in an ongoing research project (Jessie & Kendall, 2015) or in Appendix B.

The decomposition in this paper shows that the prediction of many bounded rationality models solely rely on the strategic component and is, therefore, unchanged by varying the behavioral component. However, the level of risk associated with each action is inherently built into both of these components. This is important to our main results because models that allow for agents to have a preference for risk may provide different predictions across games that have the same strategic component. With this in mind, this paper's exact decomposition would need to be modified in order to achieve the same experimental approach for models that incorporate risk (Jessie & Kendall, 2015).

This paper's decomposition is designed to normalize the component of a game concerned with personal-payoff maximization; particularly boundedly-rational maximization. Of course, this paper is not the first to recognize that humans appear to have more complex preferences than simple payoff-maximization. The previous literature in this direction has suggested that particular "social" (sometimes referred to as "other-regarding") preferences can explain non-Nash behavior. Therefore, a natural extension of this paper's results is to model subjects

that have a preference for some of these traits. The appendix carries out this extension by fitting standard models of altruism and inequity aversion in our logistic framework. In doing so, we find that models that account for altruism or inequity aversion respond to changes in the behavioral component and, therefore, produce a better fitted line than does the strict personal-payoff maximizing logit Quantal Response model.[29] Because the decomposition and subsequent experiment presented in this paper were specifically created to isolate the prediction of bounded rationality models, it is not surprising to observe models that account for social preferences fitting the data more accurately than the logit QRE. To test the predictive power of these social preference models, we then introduce a novel "corner-preference" model which has nonsensical behavioral interpretation and serves as our placebo-model test. Surprisingly, this nonsensical model has a similar predictive fit than do models that incorporate altruism or inequity-aversion. This suggests that models with an additional parameter capturing a social preference can generate a better fitted line of our experimental data. However, since we achieve similar model-fits between dissimilar and nonsensical model-motivations, this result is likely a mathematical artifact of allowing for an additional degree of freedom in the model rather than a result about any meaningful interpretation of our data.

8 Conclusion

This paper finds experimental results suggesting that humans behave differently in simple 2×2 games than is predicted by many contemporary models of bounded rationality. To do this, we apply a mathematical decomposition of games which allows us to design a laboratory experiment showing a divergence between human responsiveness in 2×2 games and a large class of bounded rationality models (Figure 2). This result suggests that subjects are influenced by changes in the game's behavioral component, which is ignored by many bounded rationality models. We offer this finding as a possible explanation for when and why models built around error-prone decision makers and/or iterative levels of sophistication fail to provide accurate predictions.

Although the primary results of our paper point to a blind spot of a widely used class of bounded rationality models, this should not be viewed in a negative way. In order to progress towards an understanding of human behavior in strategic situations, it is necessary to understand the limits of the models currently being used. By presenting the mathematical structures that are relevant (and irrelevant) to our existing models of bounded rationality, we can now understand on a global level when and for what reason these models will work (or not). This means that we can now predict whether or not one of these bounded rationality

[29] Using Akaike's Information Criterion corrected for finite samples (Akaike, 1974; Hurvich & Tsai, 1989).

models will produce a good fit of human data *before* the data is generated or collected. As discussed in section 6, this is valuable because our approach can address and unify previous findings as well as to serve as a guide for future research in behavioral economics.

Many models of bounded rationality have a general invariance that is not shared by human decision-makers. By pinpointing this disconnection, we have important information about how to proceed for future research. In particular, we now know that subjects respond to the behavioral information in a game, and the bounded rationality models do not. The obvious (but not necessarily easy) next step is to create a model that is also responsive to changes in this information, and then to generate a testable prediction for this new model, as was done here for bounded rationality models.

Appendix A: Strategically equivalent sets #2-6

$\mathcal{G}_0^2$	L	R		*strategic*			*behavioral*			*kernel*		**Data**
T	9, 11	14, 13	=	0.5, −1	$\underline{3}, \underline{1}$	+	−1.25, 1.75	1.25, 1.75	+	9.75, 10.25	9.75, 10.25	$T = .968$
B	8, 7	8, 10		−0.5, −1.5	−3, 1.5		−1.25, −1.75	1.25, −1.75		9.75, 10.25	9.75, 10.25	$L = .097$

$\mathcal{G}_{TL}^2$	L	R		*strategic*			*behavioral*			*kernel*		**Data**
T	20, 19	7, 21	=	0.5, −1	$\underline{3}, \underline{1}$	+	**7.75, 8.25**	−7.75, 8.25	+	11.75, 11.75	11.75, 11.75	$T = .968$
B	19, 2	1, 5		−0.5, −1.5	−3, 1.5		7.75, −8.25	−7.75, −8.25		11.75, 11.75	11.75, 11.75	$L = .387$

$\mathcal{G}_{TR}^2$	L	R		*strategic*			*behavioral*			*kernel*		**Data**
T	3, 18	22, 20	=	0.5, −1	$\underline{3}, \underline{1}$	+	−8.25, 7.25	**8.25, 7.25**	+	10.75, 11.75	10.75, 11.75	$T = .871$
B	2, 3	16, 6		−0.5, −1.5	−3, 1.5		−8.25, −7.25	8.25, −7.25		10.75, 11.75	10.75, 11.75	$L = .129$

$\mathcal{G}_{BL}^2$	L	R		*strategic*			*behavioral*			*kernel*		**Data**
T	19, 3	8, 5	=	0.5, −1	$\underline{3}, \underline{1}$	+	6.75, −9.75	−6.75, −9.75	+	11.75, 13.75	11.75, 13.75	$T = .645$
B	18, 22	2, 25		−0.5, −1.5	−3, 1.5		**6.75, 9.75**	−6.75, 9.75		11.75, 13.75	11.75, 13.75	$L = .258$

$\mathcal{G}_{BR}^2$	L	R		*strategic*			*behavioral*			*kernel*		**Data**
T	5, 1	27, 3	=	0.5, −1	$\underline{3}, \underline{1}$	+	−9.75, −9.25	9.75, −9.25	+	14.25, 11.25	14.25, 11.25	$T = .581$
B	4, 19	21, 22		−0.5, −1.5	−3, 1.5		−9.75, 9.25	**9.75, 9.25**		14.25, 11.25	14.25, 11.25	$L = .161$

Figure A1. Strategically equivalent set #2. Dominant strategies for both subjects. One Nash equilibrium: $(p = 1, q = 0)$.

$\mathcal{G}_0^3$	L	R		*strategic*			*behavioral*			*kernel*		**Data**
T	13, 11	10, 9	=	$\underline{2}, \underline{1}$	−1, −1	+	0, −1.25	0, −1.25	+	11, 11.25	11, 11.25	$T = .645$
B	9, 11	12, 14		−2, −1.5	$\underline{1}, \underline{1.5}$		0, 1.25	0, 1.25		11, 11.25	11, 11.25	$L = .806$

$\mathcal{G}_{TL}^3$	L	R		*strategic*			*behavioral*			*kernel*		**Data**
T	19, 20	1, 18	=	$\underline{2}, \underline{1}$	−1, −1	+	**7.5, 7.75**	−7.5, 7.75	+	9.5, 11.25	9.5, 11.25	$T = .613$
B	15, 2	3, 5		−2, −1.5	$\underline{1}, \underline{1.5}$		7.5, −7.75	−7.5, −7.75		9.5, 11.25	9.5, 11.25	$L = .645$

$\mathcal{G}_{TR}^3$	L	R		*strategic*			*behavioral*			*kernel*		**Data**
T	6, 23	20, 21	=	$\underline{2}, \underline{1}$	−1, −1	+	−8.5, 8.25	**8.5, 8.25**	+	12.5, 13.75	12.5, 13.75	$T = .742$
B	2, 4	22, 7		−2, −1.5	$\underline{1}, \underline{1.5}$		−8.5, −8.25	8.5, −8.25		12.5, 13.75	12.5, 13.75	$L = .258$

$\mathcal{G}_{BL}^3$	L	R		*strategic*			*behavioral*			*kernel*		**Data**
T	28, 3	2, 1	=	$\underline{2}, \underline{1}$	−1, −1	+	11.5, −11.25	−11.5, −11.25	+	14.5, 13.25	14.5, 13.25	$T = .355$
B	24, 23	4, 26		−2, −1.5	$\underline{1}, \underline{1.5}$		**11.5, 11.25**	−11.5, 11.25		14.5, 13.25	14.5, 13.25	$L = .742$

$\mathcal{G}_{BR}^3$	L	R		*strategic*			*behavioral*			*kernel*		**Data**
T	5, 4	23, 2	=	$\underline{2}, \underline{1}$	−1, −1	+	−10.5, −9.75	10.5, −9.75	+	13.5, 12.75	13.5, 12.75	$T = .290$
B	1, 21	25, 24		−2, −1.5	$\underline{1}, \underline{1.5}$		−10.5, 9.75	**10.5, 9.75**		13.5, 12.75	13.5, 12.75	$L = .419$

Figure A2. Strategically equivalent set #3. Battle of the sexes (coordination). Three Nash Equilibria: $(p = 1, q = 1)$, $(p = 0, q = 0)$, and $(p = \frac{3}{5}, q = \frac{1}{3})$.

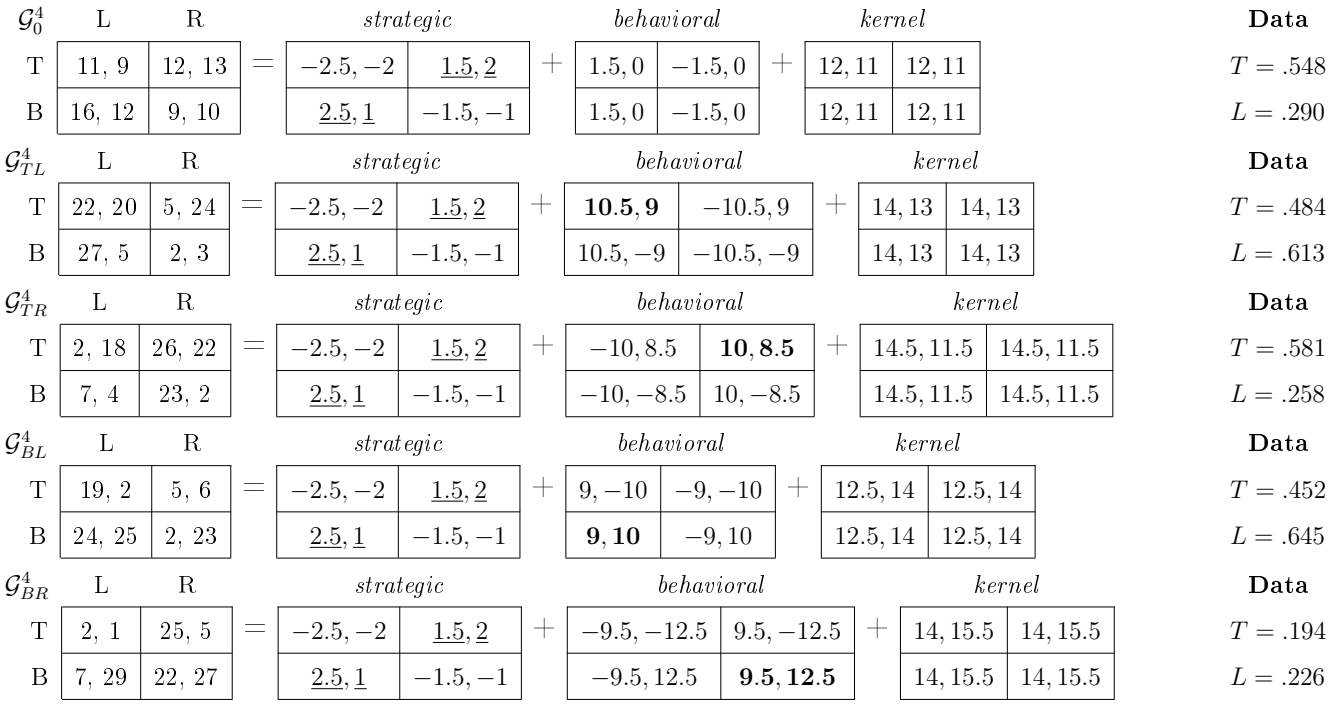

Figure A3. Strategically equivalent set #4. Battle of the sexes (anti-coordination). Three Nash Equilibria: $(p = 1, q = 0)$, $(p = 0, q = 1)$, and $(p = \frac{1}{3}, q = \frac{3}{8})$.

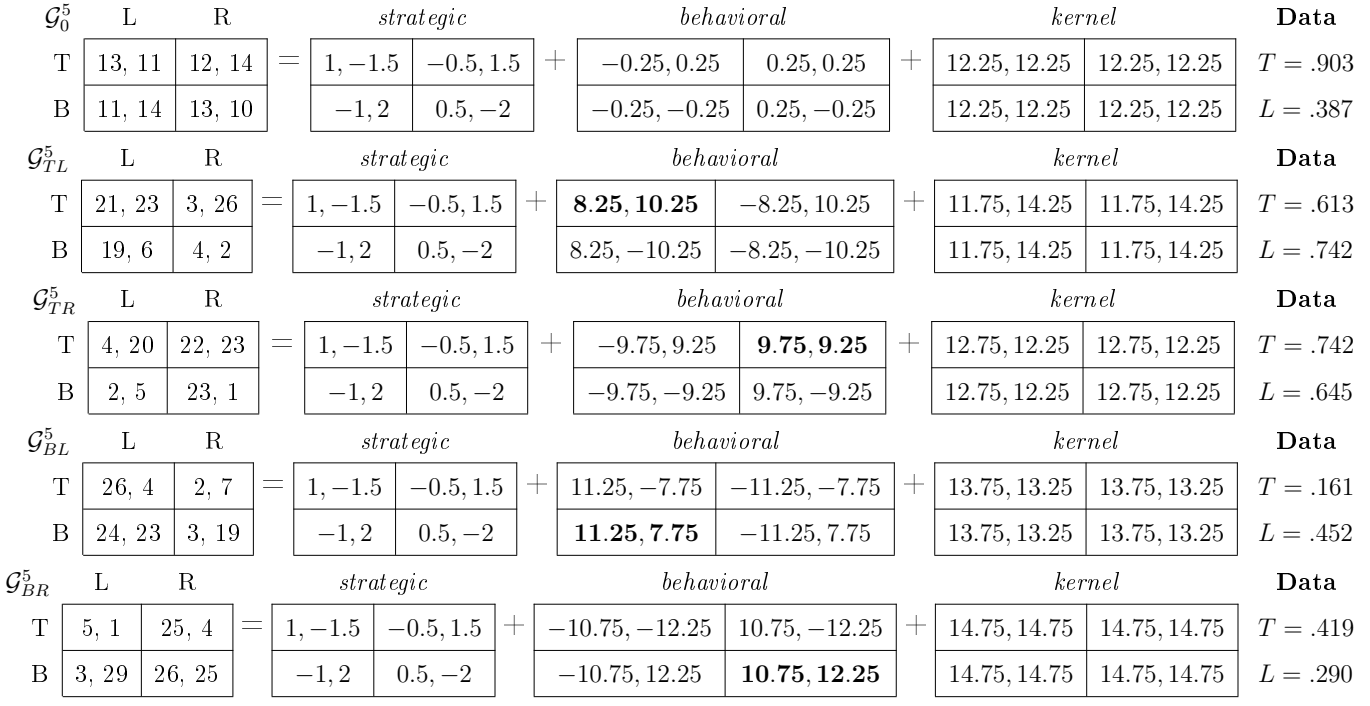

Figure A4. Strategically equivalent set #5. Matching pennies. Nash Equilibrium: $(p = \frac{4}{7}, q = \frac{1}{3})$.

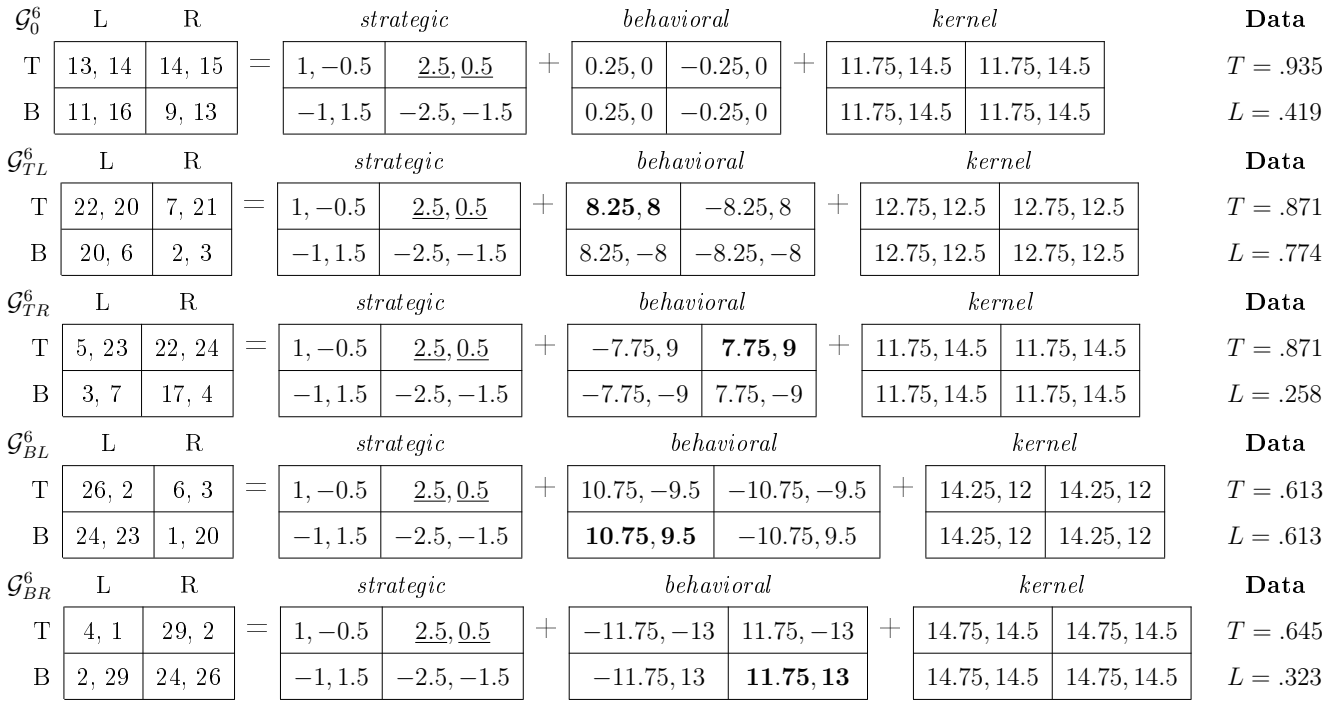

Figure A5. Strategically equivalent set #6. Dominant strategy for row subject. Nash Equilibrium: $(p = 1, q = 0)$.

Appendix B: Models of social preferences

There exist a number of different theories about human behavior that do not rely on bounded rationality, but propose instead that humans are motivated by traits other than strict personal-payoff maximization. This large literature models agents who have preferences that depend on the payoffs received by other agents. These other-regarding models have been constructed to add a human preference for fairness (Rabin, 1993 and Hoffman, McCabe, & Smith, 1996), warm-glow feelings (Andreoni, 1990), altruism (Andreoni & Miller, 2002), spitefulness (Levine, 1998), inequity aversion (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000), efficiency concerns and maximin preferences (Charness & Rabin, 2002; Engelmann & Strobel, 2004), or reciprocity (Dufwenberg & Kirchsteiger, 2004). While these concepts have been shown to accurately align with human behavior in certain environments, it has been suggested that these models are quite unstable. Small changes in framing, language, or context collapse the specified preference for others (Bénabou & Tirole, 2011). Even the bias of a subject pool has been shown to drastically change responsiveness to social preferences (Fehr, Naef, & Schmidt, 2006). Furthermore, it is likely that each model is only suited for a specific setting, thus have little predictive power about decision-making in general (Binmore & Shaked, 2010).

This section analyzes our experimental data using two well-known social preference models

of altruism and inequity aversion as well as a novel corner-preference model.[30] These social preference models have been shown to accurately fit experimental data and may explain human behavior in our experiment. This is plausible because subjects in our experiment may be motivated to maximize the payoff they receive but *in addition* they may also have a preference for how their choice affects the other subject's payoff. To explain the intuition behind these social preference models, consider Game B from the main text reproduced here.

Game B

	L	R
T	6, 5	21, 2
B	1, 24	20, 22

A personal-payoff maximizing row subject will always choose Top because there is a positive difference between 6 and 1 (if column chooses Left) as well as a positive difference between 21 and 20 (if column chooses Right). Top dominates Bottom, and a similar analysis shows that Left dominates Right. Therefore a personal-payoff maximizing subject will select Top or Left with probability 1. However, our data show that Top and Left are only chosen with probabilities .77 and .65 respectively. Here we explore three possible models of human behavior that include either the well-known preference for altruism or equity along with our novel model of "corner-bias" subject.

One explanation of the divergence between theory and observation is that human subjects also have a preference for the payoff received by the other subject. In such models, an "altruistic-row" subject will also consider how her choice affects the payoff of the column subject. This simple altruism model has been formalized many times as a utility calculation where subject i receives a payoff discounted by μ for choosing an action that increases subject j's payoff.

$$U_i = \pi_i + \mu\pi_j \tag{11}$$

Consider altruistic-row's decision in Game B given that the column subject chooses Left. Altruistic-row's personal payoff is greater by choosing Top by a factor of 5 (6-1), but this is counterbalanced by the column subject becoming 19 worse off (5-24) by altruistic-row's decision to choose Top instead of Bottom. A similar story can be told given that the column subject chooses Right. If the altruistic-row subject is sufficiently motivated by the payoff rewarded to the column subject (if μ is large enough), then she will choose Bottom. Each subject's choice is dependent on their preference for altruism, μ. In Game B, the altruistic-row subject will choose Top if the following inequality holds, and Bottom otherwise.

[30]The analysis in this model focuses on the row subject. The same qualitative results are true for the column subject.

$$E_{Row}\pi(T) > E_{Row}\pi(B)$$
$$q \cdot (\pi_{Row,TL} + \mu \cdot \pi_{Col,TL}) + (1-q) \cdot (\pi_{Row,TR} + \mu \cdot \pi_{Col,TR}) >$$
$$q \cdot (\pi_{Row,BL} + \mu \cdot \pi_{Col,BL}) + (1-q) \cdot (\pi_{Row,BR} + \mu \cdot \pi_{Col,BR})$$
$$q \cdot (6 + \mu \cdot 5) + (1-q) \cdot (21 + \mu \cdot 2) > q \cdot (1 + \mu \cdot 24) + (1-q) \cdot (20 + \mu \cdot 22) \quad (12)$$

We model these altruistic subjects using a similar logit-choice model of decision-making used in the main paper. This generates a model with two estimable parameters that is fitted to our experimental data for the row subject in Figure B1.[31] The Akaike Information Criterion corrected for finite sample size suggests that the altruistic model fits the data far better than the logit QRE: AICc(Altruism) =75.92<90.04 =AICc(logit QRE).[32]

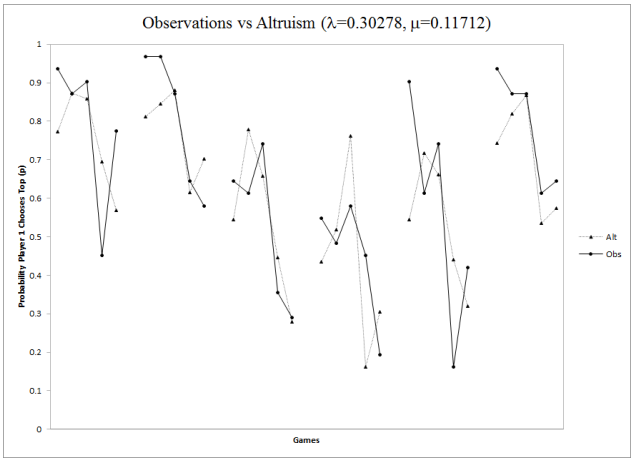

Figure B1. This figure shows the row subject's aggregated observed strategy choices in each of the 30 games along with the estimated fit using the logit QRE with altruistic subjects.

Figure B1 suggests that subjects in the lab may have been motivated by altruistic preferences. In order to test this qualitative finding, we also tested a common model that includes a subject's preference for equity of payoffs, rather than for altruism. Such an "egalitarian-row" subject will have a preference for how her choice affects the difference between her payoff and the payoff of the column subject. Similar to the altruism model, this model of inequity-aversion has been formalized in a utility calculation where an subject receives a payoff discounted by either μ_1 or μ_2 depending on whether or not the inequity is in her respective favor.

$$U_i = \pi_i + \mu_1 \cdot Max\{\pi_i^1 - \pi_j^1, 0\} + \mu_2 \cdot Max\{\pi_i^2 - \pi_j^2, 0\} \quad (13)$$

[31] This model combining noisy decisions and altruism is similar to the one used in Goeree, Holt, and Laury, 2002.

[32] The interpretation of these test statistics is that the logit QRE model is exp((75.92=90.04)/2) = .0009 times as probable to minimize the information loss as the altruism model.

Consider an egalitarian-row subject's decision to choose Top in game B given that the column subject chooses Left. Egalitarian-row would take into account that her personal payoff is greater by choosing Top by a factor of 5 (6-1). In addition, Top reduces the inequity between the two agents because the difference between 1 and 24 is greater than the difference between 6 and 5. Also notice that choosing Top or Bottom will determine which agent will be favored by the inequity. A similar comparison is used given that the column subject chooses Right. The choice made by egalitarian-row may depend on how strong her preference is for the equity of payoffs. In Game B, the egalitarian-row subject will choose Top if the following inequality holds, and Bottom otherwise.

$$E_{Row}\pi(T) > E_{Row}\pi(B)$$

$$q \cdot (\pi_{Row,TL} + \mu_1 \cdot Max\{\pi_{Row,TL} - \pi_{Col,TL}, 0\} + \mu_2 \cdot Max\{\pi_{Col,TL} - \pi_{Row,TL}, 0\}) + (1-q) \cdot (\pi_{Row,TR} + \mu_1 \cdot Max\{\pi_{Row,TR} - \pi_{Col,TR}, 0\} + \mu_2 \cdot Max\{\pi_{Col,TR} - \pi_{Row,TR}, 0\}) > q \cdot (\pi_{Row,BL} + \mu_1 \cdot Max\{\pi_{Row,BL} - \pi_{Col,BL}, 0\} + \mu_2 \cdot Max\{\pi_{Col,BL} - \pi_{Row,BL}, 0\}) + (1-q) \cdot (\pi_{Row,BR} + \mu_1 \cdot Max\{\pi_{Row,BR} - \pi_{Col,BR}, 0\} + \mu_2 \cdot Max\{\pi_{Col,BR} - \pi_{Row,BR}, 0\})$$

$$q \cdot (6 + \mu_1 \cdot Max\{6-5, 0\} + \mu_2 \cdot Max\{5-6, 0\})) + (1-q) \cdot (21 + \mu_1 \cdot Max\{21-2, 0\} + \mu_2 \cdot Max\{2-21, 0\})) > q \cdot (1 + \mu_1 \cdot Max\{1-24, 0\} + \mu_2 \cdot Max\{24-1, 0\})) + (1-q) \cdot (20 + \mu_1 \cdot Max\{20-22, 0\} + \mu_2 \cdot Max\{22-20, 0\}))$$

$$q \cdot (6 + \mu_1 \cdot 1) + (1-q) \cdot (21 + \mu_1 \cdot 19) > q \cdot (1 + \mu_2 \cdot 23) + (1-q) \cdot (20 + \mu_2 \cdot 2) \qquad (14)$$

We use the same logit-choice model of decision-making to model these inequity-averse subject. This generates a model with three estimable parameters that is fit to our experimental data in Figure B2.[33]

[33]This model combining noisy decisions and inequity aversion is used in Blanco, Engelmann, and Normann, 2011.

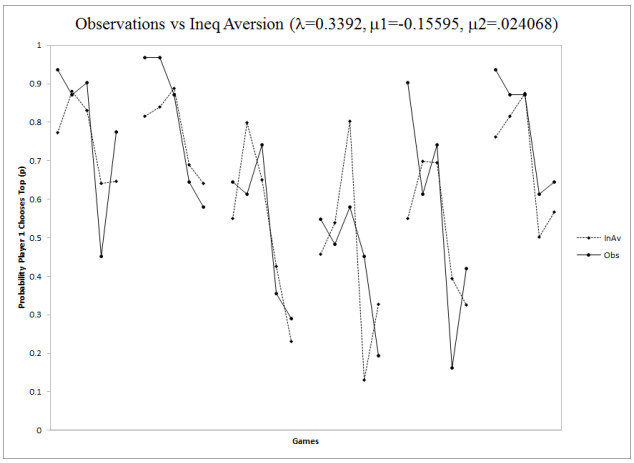

Figure B2. This figure shows the row subject's aggregated observed strategy choices in each of the 30 games along with the estimated fit using the logit QRE with inequity-averse subjects.

A model including a preference for equity fits the data with equal success as a model including altruism. The observed data, altruism model, and inequity-aversion model are shown in Figure B3. The AICc suggests that the altruism model fits the data slightly better than the inequity-aversion model: AICc(Altruism) =75.92<78.16 =AICc(Inequity-aversion).[34]

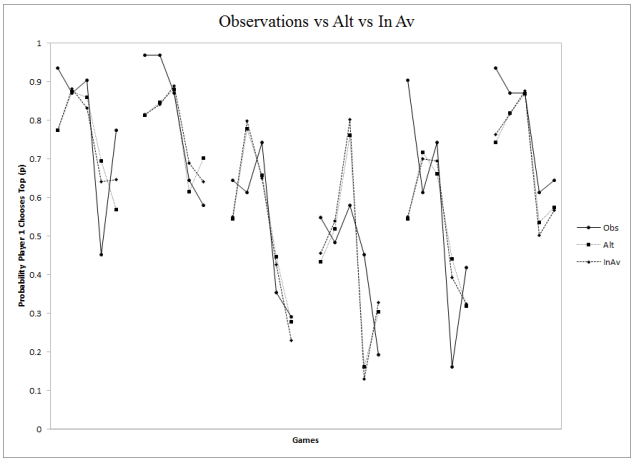

Figure B3. This figure shows the row subject's aggregated observed strategy choices in each of the 30 games along with the estimated fit using the logit QRE with altruistic and inequity-averse subjects.

Finally, consider a different model introduced here where subjects have a preference for the corner payoffs in a 2×2 game. In this new corner model, subjects have a preference over the payoff differences between their payoff in a cell and the other subject's payoff in

[34]The interpretation of these test statistics is that the inequity-aversion model is exp((75.92=78.16)/2) = .3263 times as probable to minimize the information loss as the altruism model.

the opposite corner-cell of the 2×2 matrix. We purposefully constructed this model so that there is no plausible reason to suspect that humans would have such a preference. However, we can formalize this corner model in a utility calculation similar to the altruism model and inequity-aversion model.

$$U_i = \pi_i + \mu \pi_{corner} \tag{15}$$

Consider corner-row subject's decision to choose Top in game B given that the column subject chooses Left. As before, corner-row's personal payoff is greater by choosing Top by a factor of 5 (6-1). However, this is counterbalanced by considering the difference between the 6 earned by the corner-row subject in the Top-Left corner cell and the 22 earned by the column subject in the Bottom-Right corner cell. A similar comparison is used given that the column subject chooses Right. The choice made by the corner-row subject depends on how strong her preference is over the difference in the payoffs of the corner cells (μ). In Game B, the corner-row subject will choose Top if the following inequality holds, and Bottom otherwise.

$$\begin{gathered} E_{Row}\pi(T) > E_{Row}\pi(B) \\ q \cdot (\pi_{Row,TL} + \mu \cdot \pi_{Row,TL}) + (1-q) \cdot (\pi_{Row,TR} + \mu \cdot \pi_{Row,TR}) > \\ q \cdot (\pi_{Row,BL} + \mu \cdot \pi_{Col,BR}) + (1-q) \cdot (\pi_{Row,BR} + \mu \cdot \pi_{Col,BL}) \\ q \cdot (6 + \mu \cdot 6) + (1-q) \cdot (21 + \mu \cdot 21) > q \cdot (1 + \mu \cdot 22) + (1-q) \cdot (20 + \mu \cdot 24) \end{gathered} \tag{16}$$

Of course, the comparison between one subject's payoff in a cell with the other subject's payoff in the corner cell should have no influence on an agent's choice. There is no reasonable story suggesting why a human would compare these numbers, and the resulting comparison for the row subject's choice is nonsensical (as is Column agent's choice). Because this model lacks a convincing motivation, we should expect that a model of corner-bias agents will produce a poor prediction of human behavior in our experiment. This model is formalized similarly to the altruistic and inequity-averse models with a logistic error structure and a two-parameter estimable model with μ measuring preference for corner-bias. This model is fit to our experimental data in Figure B4.

The model of corner-bias agents fits the data surprisingly well. In fact, the AICc suggests that the corner model fit the data slightly better than the inequity-aversion model: AICc(Corner) =77.99<78.16 =AICc(Inequity-aversion). Even though the corner model is slightly outperformed by the altruism model, as Figure B5 illustrates, both models have resoundingly similar predictions.

Fitting models to our experimental data suggests that subjects in our experiment appear to be equally as altruistic as they are inequity averse as they are corner-bias. Of course, we hesitate to promote that any of these are the correct behavioral interpretation of human

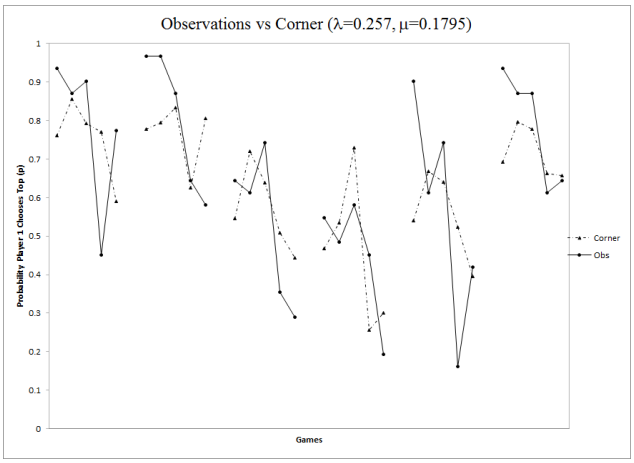

Figure B4. This figure shows the row subject's aggregated observed strategy choices in each of the 30 games along with the estimated fit using the logit QRE with corner-bias subjects.

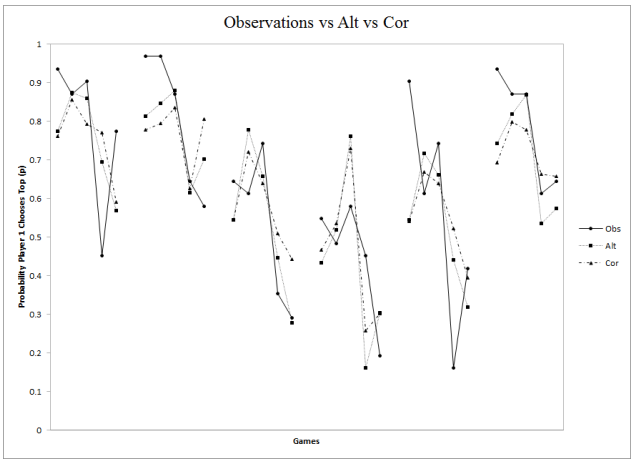

Figure B5. This figure shows the row subject's aggregated observed strategy choices in each of the 30 games along with the estimated fit using the logit QRE with altruistic subjects and corner-bias subjects.

preferences (especially the nonsensical corner model). However, the fact that the corner model fits as well as the two traditional models is concerning for the meaningful interpretation of any of these models to our data.

References

[1] **Akaike, Hirotugu**. 1974. "A new look at the statistical model identification." *Automatic Control, IEEE Transactions on* 19(6): 716-723.

[2] **Andreoni, James**. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm Glow Giving." *The Economic Journal*, 100: 464 - 477.

[3] **Andreoni, James and John Miller**. 2002. "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica*, 70(2): 737 - 753.

[4] **Bénabou, Roland and Jean Tirole**. 2011. "Identity, Morals, and Taboos: Beliefs as Assets." *The Quarterly Journal of Economics*, 126(2): 805-855.

[5] **Binmore, Ken and Avner Shaked**. 2010. "Experimental Economics: Where Next?" *Journal of Economics Behavior & Organization*, 73: 87 - 100.

[6] **Bolton, Gary E. and Axel Ockenfels**. 2000 "ERC: A theory of equity, reciprocity, and competition." *American Economic Review*, 166-193.

[7] **Blanco, Mariana, Dirk Engelmann, and Hans Theo Normann**. 2011. "A within-subject analysis of other-regarding preferences." *Games and Economic Behavior*, 72(2): 321-338.

[8] **Camerer, Colin F., Teck-Hua Ho, and Juin-Kuan Chong**. 2004. "A Cognitive Hierarchy Model of Games." *The Quarterly Journal of Economics*, 119(3), 861-898.

[9] **Charness, Gary and Matthew Rabin.** 2002. "Understanding social preferences with simple tests." *Quarterly Journal of Economics*, 817-869.

[10] **Crawford, Vincent P., Miguel A. Costa-Gomes, and Nagore Iriberri**. 2013. "Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications." *Journal of Economic Literature*, 51(1), 5-62.

[11] **Deck, Cary A**. 2001. "A test of game-theoretic and behavioral models of play in exchange and insurance environments." *American Economic Review*, 1546-1555.

[12] **Dufwenberg, Martin and Georg Kirchsteiger**. 2004. "A Theory of Sequential Reciprocity" *Games and Economic Behavior*, 47: 268 - 298.

[13] **Engelmann, Dirk and Martin Strobel**. 2004. "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments." *American Economic Review*, 857-869.

[14] **Fehr, Ernst and Klaus M. Schmidt**. 1999. "A Theory of Fairness, Competition, and Cooperation" *Quarterly Journal of Economics*, 114(3): 817 - 868.

[15] **Fehr, Ernst, Michael Naef, and Klaus M. Schmidt**. 2006. "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments: Comment." *American Economic Review*, 96(5): 1912-1917.

[16] **Fischbacher, Urs**. 2007. "Z-Tree: Zurich Toolbox for Read-made Economic Experiments" *Experimental Economics*, 10: 171-178.

[17] **Goeree, Jacob and Charles Holt**. 2004. "A Model of Noisy Introspection" *Games and Economic Behavior*, 46(2): 365 - 382.

[18] **Goeree, Jacob and Charles Holt**. 2005. "An Experimental Study of Costly Coordination" *Games and Economic Behavior*, 46(2): 281 - 294.

[19] **Goeree, Jacob, Charles Holt, and Susan Laury**. 2002. "Private costs and public benefits: unraveling the effects of altruism and noisy behavior." *Journal of Public Economics*, 83(2): 255-276.

[20] **Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith**. 1996. "Social Distance and Other Regarding Behavior in Dictator Games" *American Economic Review*, 86(3): 653-660.

[21] **Hurvich, Clifford M. and Chih-Ling Tsai**. 1989. "Regression and time series model selection in small samples." *Biometrika* 76(2): 297-307.

[22] **Jessie, Daniel and Ryan Kendall**. 2015. "Decomposing models of risk" [Working Paper].

[23] **Jessie, Daniel and Donald G. Saari**. 2013. "Strategic and Behavioral Decomposition of $2 \times 2 \times \cdots \times 2$ Games," Technical Report, April 2013, Institute for Mathematical Behavioral Sciences, University of California, Irvine.

[24] **Levine, David K.**. 1998. "Modeling Altruism and Spitefulness in Experiments" *Review of Economic Dynamics*, 1: 593 - 622.

[25] **Levine, David K. and Thomas R. Palfrey**. 2007. "The Paradox of Voter Participation? A Laboratory Study" *The American Political Science Review*, 101(1): 143 - 158.

[26] **Luce, Duncan**. 1959. *Individual Choice Behavior.* New York, Wesley.

[27] **McFadden, Daniel**. 1973 *Conditional logit analysis of qualitative choice behavior.*

[28] **McKelvey, Richard D. and Thomas R. Palfrey**. 1995. "Quantal Response Equilibrium for Normal Form Games" *Games and Economic Behavior*, 10: 6 - 38.

[29] **McKelvey, Richard D., Thomas R. Palfrey, and Roberto A. Weber**. 2000. "The effects of payoff magnitude and heterogeneity on behavior in 2×2 games with unique mixed strategy equilibria." *Journal of Economic Behavior & Organization*, 42(4): 523-548.

[30] **Nagel, Rosemarie**. 1995. "Unraveling in guessing games: An experimental study." *American Economic Review*, 85(5): 1313-1326.

[31] **Nash, John F.**. 1950. "Equilibrium Points in N-Person Games" *Proceedings of the National Academy of Sciences*, 36: 48 - 49.

[32] **Nash, John F.**. 1951. "Non-Cooperative Games" *Annals of Mathematics*, 2(54): 286 - 295.

[33] **Rabin, Matthew**. 1993. "Endogenous Preferences in Games" *American Economics Review*, 83: 1281 - 1302.

[34] **Rogers, Brian W., Thomas R. Palfrey, and Colin F. Camerer**. 2009. "Heterogeneous quantal response equilibrium and cognitive hierarchies." *Journal of Economic Theory*, 144(4), 1440-1467.

[35] **Selten, Reinhard and Thorsten Chmura**. 2008. "Stationary Concepts for Experimental 2×2 Games" *American Economics Review*, 98(3): 938 - 966.

[36] **Stahl, Dale O. and Paul W. Wilson**. 1994. "Experimental evidence on players' models of other players." *Journal of Economic Behavior & Organization*, 25(3): 309-327.

[37] **Weizsäcker, Georg**. 2003. "Ignoring the rationality of others: evidence from experimental normal-form games." *Games and Economic Behavior*, 44(1): 145-171.