Identifying Predictable Players:[*] Relating Behavioral Types and Subjects with Deterministic Rules

Daniel E. Fragiadakis *Stanford University*

Daniel T. Knoepfle *Stanford University*

Muriel Niederle *Stanford University* and *NBER*

December 29, 2013

Abstract

Behavioral game theory models are important in organizing experimental data of strategic decision making. However, are subjects classified as behavioral types more predictable than unclassified subjects? Alternatively, how many predictable subjects await new behavioral models to describe them? In our experiments, subjects play simple guessing games against random opponents and are subsequently asked to replicate or best-respond to their past choices. We find that existing behavioral game theory types capture 2/3 of strategic subjects, i.e., individuals who can best respond. However, there is additional room for non-strategic rule-of-thumb models to describe subjects who can merely replicate their actions.

1 Introduction

A robust finding of strategic choice experiments is that deviations from Nash equilibrium are common. This has lead to alternative behavioral models with varying specifications of beliefs and derived choices; of these, hierarchy models, particularly the level-k model, are probably the most prominent.[1] In a typical empirical paper, participants in the laboratory play a

[*] We are especially grateful to Asen Ivanov for his impact on the design of the experiment. We thank Vince Crawford, Guillaume Fréchette and Matt Jackson for helpful comments and the NSF for generous support.

[1] A level-k player best-responds to beliefs that opponents are level-$(k-1)$, with a level-0 player assumed to randomly choose any action or to choose a fixed action considered to be focal. The model originated in empirical papers that found it rationalized large fractions of behavior in beauty contest games (Nagel, 1995) and small normal-form games (Stahl and Wilson, 1994, 1995). The level-k model has since been used to model strategic behavior in a multitude of experiments, and has spawned a literature on extensions and theoretical underpinnings. A notable variant is the cognitive hierarchy model (Camerer, Ho, and Chong, 2004), in which frequencies of types k in the population are assumed to be distributed according to some distribution, and a player of type k has beliefs about opponent types corresponding to this distribution truncated at $k-1$.

set of games and are then classified as specific behavioral types given their observed choices; see Crawford, Costa-Gomes, and Iriberri (2013) for an overview.[2] Such studies generally allow some discrepancy between subjects' choices and the models, with subjects assumed to be implementing their behavioral types' prescriptions with error. Beyond some threshold, subjects are left unclassified and unexplained.

This literature leads naturally to two questions that we address in this paper. The first concerns the extent to which existing behavioral game theory types coincide with the set of participants who play deliberately according to deterministic rules. We construct an environment and a test allowing us to assess whether subjects use a deterministic rule, even for subjects whose rules are unknown to us. If existing behavioral game theory models capture most subjects who deliberately use a deterministic rule, then we expect subjects matching behavioral game theory types to score better on our test than other subjects. Other well performing subjects not matching existing models indicate there is deliberate behavior that may be captured by future models. The second question concerns the boundaries of classification methods given an existing set of behavioral types. We provide a quantification of the downside of classifying a greater number of subjects.

In our experiments, subjects first play a sequence of 20 two-player guessing games (of the form in Costa-Gomes and Crawford, 2006, henceforth CGC) with anonymous opponents and without feedback. We then present each subject a series of strategic decisions dependent on her original choices. These second-phase choices are unanticipated and we will show that exact purely numeric memory of Phase I choices is very limited. Our two main treatments differ only in their Phase II design.

In Phase II of the *Replicate* treatment, a subject plays the same 20 games in the same order and player role as in Phase I. However, subjects are now paid as a function of how close their Phase II guesses are to their corresponding Phase I guesses. Any subject who deliberately uses a well-defined deterministic rule and is aware of doing so should be able to replicate her actions. Given reasonable assumptions of self-awareness and cognitive ability, a failure to replicate one's actions suggests substantial idiosyncratic randomness in decisions.

In Phase II of the *BestRespond* treatment, participants play the same 20 games once more in the same order; however, they now take the role previously occupied by their opponents. The subject is informed she plays against a computer whose guess in each game is the guess she herself made in that game in Phase I; however, she is not reminded of her exact Phase I choices. In effect, subjects are playing against their own first-phase behavior.[3] The payoff-maximizing

[2] In addition to the works mentioned above, some leading examples of papers that seek to classify participants are Costa-Gomes, Crawford, and Broseta (2001) for normal form games, and Crawford, Gneezy, and Rottenstreich (2008) for coordination games.

[3] This treatment is inspired by the design in Ivanov, Levin, and Niederle (2010) but has important differences

choice is the best-response in that game to the subject's original Phase I action. We reason that any subject who is playing purely according to a deliberate rule in Phase I, is aware of doing so, and who best-responds to beliefs about her opponents' play, should be able to first replicate her former guess and then find the best-response to it.

Our results show most subjects are unsuccessful at replicating or best-responding to their past behavior; in both treatments, fewer than half of the participants exceed a permissive 40% threshold for the number of optimal Phase II choices. This aggregate failure suggests that much of the observed behavior is idiosyncratic, even to the decision-makers themselves. However, such a pessimistic outlook for the overall scope of behavioral game theory is counterbalanced by the success of its existing models in explaining the deliberate subjects.

Using a conventional approach and the Phase I observations, we classify subjects as matching given behavioral types (from a set of models that includes equilibrium, level-k, and others). Subjects whose behavior does not match better than a specific threshold are left unclassified.

We find that a vast majority of subjects classified as matching a behavioral type are able to replicate and best-respond to their past behavior. Furthermore, classified subjects are substantially more successful in replicating or best-responding to their former guesses than are subjects not classified as a behavioral type. These results unequivocally confirm the success of existing behavioral models like level-k in accurately describing the decision-making of deliberate subjects. The remaining subjects as a group are different and cannot be described as equally deliberate decision makers who use well-defined rules in a similar manner.

In addition, our environment allows us to distinguish between strategic behavior and systematic but non-strategic behavior. For instance, while the level-k model is often described as best-response to non-equilibrium beliefs, some researchers have argued that the lowest levels, such as $L1$, may instead arise as players using "rules of thumb", that is behaving in systematic but not very strategic ways. Our *BestRespond* treatment is more strategically demanding than the *Replicate* treatment, requiring a change in behavior in response to the change in opponent. As the additional strategic reasoning required is minimal, the distinction might appear trivial; however, our empirical results show it to be significant. We show that classified subjects can best-respond to their former guesses just as well as they can replicate them. In contrast, subjects unclassified fail to best-respond to their past behavior much more so than they fail to replicate it. Subjects who can replicate their guesses but fail to best-respond to them are probably better described as using rules of thumb than as best-responding to beliefs.

Overall, only about 40% of subjects who have high rates of replicating their former guesses are classified as behavioral types, suggesting appreciable room for additional behavioral game

we discuss below. A design in which subjects play against themselves is also a central component of Blume and Gneezy (2010).

theory models. However, existing behavioral models have been very successful in identifying a large majority of strategic subjects. Of subjects who have high rates of best-responding to their former guesses, roughly two-thirds are behavioral types. Note that Nash equilibrium only accounts for about twenty-six percent of strategic subjects, while the level-k model accounts for an additional 35%.[4] There is still room for behavioral game models to describe the remaining 32% of strategic subjects.

In the second part of the paper we confine attention to an existing set of behavioral game theory types. In addition to our previous approach we now use a maximum likelihood estimation to classify subjects. As we relax classification criteria more subjects are classified as behavioral types. Our design allows us to show the trade-off between classifying more subjects and capturing subjects whose Phase II type matches that implied by their Phase I classification.

The paper proceeds as follows: Section 2 describes the experiment and Section 3 the non-parametric classification of subjects. In Section 4 we present results pertaining to the predictability of classified and non-classified subjects using a model-free approach. Section 5 confines attention to behavioral game theory types and discusses the trade-off between classifying more subjects and capturing subjects whose behavior conforms to the behavior predicted by their Phase I play. Section 6 discusses the literature and we conclude in Section 7.

2 The Experiment

2.1 Two-Person Guessing Games

Participants interact in simple complete information "two-person guessing games".[5] In a two-person guessing game, player i facing opponent j wishes to guess as close as possible to her goal, which equals her target multiple t_i times her opponent's guess x_j. Likewise, player j's goal equals his target multiple t_j times x_i. Each player has a range of allowable guesses $[l_i, u_i]$, and the two players simultaneously submit guesses x_i and x_j. The payoff of i is a function of the realized distance from the player's guess x_i to her goal $t_i x_j$, $e_i = |x_i - t_i x_j|$; the function strictly decreases until the payoff reaches zero. We present the 20 games used in the experiment, as well as the predictions of various behavioral game theory models in Table 1. Further details are given later in this section.

[4] The dominance-k model adds another 6%.

[5] Another "two-person guessing game" is that of Grosskopf and Nagel (2008). They consider the familiar "p-beauty contest" guessing game where n players guess a number between 0 and 100, and the winner is the player closest to p times the mean of all submitted guesses, with $p < 1$. When $n = 2$, as in their experiments, guessing 0 becomes a dominant strategy. We opt for CGC games as they allow us to have subjects play many different games in which models that have agents best-respond to beliefs result in different actions.

	Lower Limit	**Upper Limit**	**Target**
DM1 (YOU)	l_1	u_1	t_1
DM2 (OTHER PARTICIPANT)	l_2	u_2	t_2

FIGURE 1.—Presentation of game parameters in Phase I

2.2 Experimental Treatments

All treatments but one shared a common two-phase structure. In Phase I, subjects played a series of 20 two-person guessing games against anonymous opponents without feedback. Game parameters were public information in all games and were presented as in Figure 1. In Phase II, subjects were tasked with either replicating or best-responding to their own first-phase choices in the same series of games.

The experiment consisted of the *Replicate*, *BestRespond*, *ShowGuesses*, and *Memory* treatments. The Phase I tasks of the *Replicate*, *BestRespond*, and *ShowGuesses* treatments were the same and are described in a single subsection below. We then explain Phase II of each of these treatments separately. Finally, we discuss the *Memory* treatment.

2.2.1 Phase I of the *Replicate*, *BestRespond*, and *ShowGuesses* Treatments

Subjects played all 20 games in individually-specific random orders without feedback on realized payoffs or opponents' guesses. Subjects were randomly and anonymously rematched with opponents before each game.[6] Subjects always saw themselves in the role of player 1 (called "Decision Maker 1" or "DM1") in instructions and the experimental task, as shown in Figure 1. If i was matched to opponent j in a given trial, she wished to make a guess x_i as close as possible to her goal $t_i x_j$ and earned a payoff decreasing in $e_i = |x_i - t_i x_j|$.

2.2.2 Phase II of the *Replicate* Treatment

In Phase II of the *Replicate* treatment, a subject faced the same sequence of 20 games from Phase I in the same order.[7] Participants were told they would be paid as a function of how close their Phase II guess was to the guess they made in Phase I in the same game. The payoff

[6] Unknown to subjects, in each session we split them into two equal-sized groups, G1 and G2. The group specified the game role in each of the 20 games; that is, the matching was such that each pair of opposing players consisted of one member of G1 and one member of G2.

[7] The motivation behind the preservation of order across phases was twofold. First, subjects may switch rules during Phase I and only remember the *number* of games played before the switch; they may not remember the specific games for which they used each of their rules. Second, for every game in Phase I, the subject made the same number of guesses, 19, before seeing that same game again in Phase II. On average, for both this treatment as well as the *BestRespond* treatment, 45 minutes have passed between playing a given game in Phase I and playing that same game in Phase II.

Table 1.—The 20 two-person guessing games and behavioral game theory type guesses

	ts	player	l	u	t	L1	L2	L3	EQ	D1	D2	Game
Crawford and Costa-Gomes Games	strong	G1	100	900	0.5	150	250	112.5	100	162.5	131.25	1
		G2	100	500	1.5	500	225	375	150	262.5	262.5	
		G1	300	900	0.7	350	546	318.5	300	451.5	423.15	2
		G2	100	900	1.3	780	455	709.8	390	604.5	604.5	
		G1, G2	100	500	0.7	210	315	220.5	350	227.5	227.5	3
		G1, G2	100	500	1.5	450	315	472.5	500	337.5	341.25	4
		G1, G2	300	500	0.7	350	420	367.5	500	420	420	5
		G1, G2	100	900	1.5	600	525	630	750	600	611.25	6
	weak	G1	300	900	1.3	780	900	900	900	838.5	900	7
		G2	300	900	1.3	780	900	900	900	838.5	900	
		G1	300	500	1.5	500	500	500	500	500	500	8
		G2	300	900	1.3	520	650	650	650	617.5	650	
		G1	100	500	0.7	350	105	122.5	100	122.5	122.5	9
		G2	100	900	0.5	150	175	100	100	150	100	
		G1	100	900	0.5	200	175	150	150	200	150	10
		G2	300	500	0.7	350	300	300	300	300	300	

	ts	player	l	u	t	L1	L2	L3	EQ	D1	D2	Game
Generated Games	strong	G1	250	500	1.1	500	330	440	250	330	330	11
		G2	150	950	0.8	300	400	264	200	300	276	
		G1	100	750	0.8	400	510	480	750	440	440	12
		G2	50	950	1.5	637.5	600	765	950	637.5	652.5	
		G1	150	750	1.5	712.5	337.5	534.38	150	337.5	337.5	13
		G2	50	900	0.5	225	356.25	168.75	75	225	178.12	
		G1	200	800	1.3	617.5	455	561.92	200	455	455	14
		G2	50	900	0.7	350	432.25	318.5	140	350	324.8	
		G1, G2	250	1000	1.5	937.5	562.5	843.75	375	637.5	637.5	15
		G1, G2	250	1000	0.6	375	562.5	337.5	250	412.5	382.5	16
		G1, G2	100	950	0.5	225	375	168.75	100	225	178.12	17
		G1, G2	150	750	1.5	750	337.5	562.5	150	356.25	356.25	18
	weak	G1	350	500	0.5	350	350	350	350	350	350	19
		G2	450	700	1.3	552.5	455	455	455	455	455	
		G1	450	850	1.1	467.5	660	660	660	676.5	676.5	20
		G2	250	600	1.4	600	600	600	600	600	600	

The game parameters $\{[l_1, u_1], t_1; [l_2, u_2], t_2\}$ and model predictions for all 20 games are reported above. The player, G1 or G2, specifies the role taken by the group's players in that game in Phase I. For each game we also give the source of the game (from CGC or generated by us) and the quality of type separation (ts), strong or weak, where strong type separation requires that $L1$, $L2$, $L3$, $L4$, and EQ are separated by at least 10 units. Each subject played each game once, where games 3, 4, 5, 6, 15, 16, 17 and 18 are played from both sides.

as a function of the distance from the goal was identical to Phase I. More precisely, for a given game, let x_i^I be the guess subject i made in Phase I and x_i^{II} be her guess in Phase II. Then subject i's payoff from the Phase II decision was a decreasing function of $e_i = |x_i^{II} - x_i^I|$.

2.2.3 Phase II of the *BestRespond* Treatment

In Phase II of the *BestRespond* treatment, a subject faced the same sequence of 20 games from Phase I in exactly the same order. Now, however, subjects were informed they would play in the role of player 2, while the role of player 1 they had occupied in Phase I would be taken by the computer. Subjects were told the computer would make the exact same guess that the subject previously made when playing the game in Phase I. Therefore, if subject i made a guess of x_i^I in Phase I in game $\{[l_1, u_1], t_1; [l_2, u_2], t_2\}$, then she wished to make a guess x_i^{II} in Phase II as close as possible to $t_2 x_i^I$. The payoff as a function of the distance from the goal, $e_i = |x_i^{II} - t_2 x_i^I|$, was identical to Phase I. Subjects were not shown their previous Phase I guesses when making their Phase II guess. The games were presented to subjects as in Figure 2.

To maximize earnings in the *BestRespond* treatment, subjects have to understand that the information that player 1 is replaced by the computer that uses the subjects' Phase I choices is valuable in fixing their beliefs on the opponents' behavior. Using this insight, they need to first replicate their own former guess, and then compute the best-response.

The need for the *BestRespond* treatment arises from the fact that it is not obvious that subjects who succeed in replicating their Phase I guesses do so because those guesses resulted from deliberate strategic choice. For example, it has been argued that the level-k model may merely coincide with non-strategic rules of thumb, especially for low level-k types like $L1$ (see e.g. Coricelli and Nagel, 2009 and Crawford, Costa-Gomes, and Iriberri, 2013). A subject i who uses a rule of thumb may not recognize the value of the information that the action of the opponent in Phase II is i's Phase I action. We therefore expect a subject whose behavior derives from a non-strategic rule of thumb to be able to replicate her past behavior, but not necessarily best-respond to it.[8]

2.2.4 Control: Phase II of the *ShowGuesses* Treatment

Failure to best-respond in Phase II of the *BestRespond* treatment could also simply derive from difficulty in understanding or willingness in executing the computations necessary to determine

[8] Cooper and Kagel (2005) provide compelling evidence that subjects often fail to play strategically because they fail to think about the behavior of their opponent. However, it could be that some subjects whose initial behavior was produced by a rule of thumb were able to best-respond to that behavior, as success in Phase II of the *BestRespond* treatment requires only a minimal form of strategic thinking.

	Lower Limit	**Upper Limit**	**Target**
DM2 (YOU)	l_2	u_2	t_2
DM1 (COMPUTER)	l_1	u_1	t_1

FIGURE 2.—Presentation of game parameters in Phase II of the BestRespond treatment

the best response to a guess. The *ShowGuesses* treatment provides a control for this hypothesis. Phase II of the *ShowGuesses* treatment is identical to Phase II of the *BestRespond* treatment with one exception: a subject is shown her Phase I guess when prompted for her Phase II guess. Being able to best-respond to a shown guess seems a minimal requirement for subjects who make deliberate strategic choices, that is, subjects who form potentially non-equilibrium beliefs about the behavior of their opponents and best-respond to these beliefs.

2.2.5 Control: *Memory* Treatment

We presume that a subject who successfully replicates or best-responds to her past guesses in our main treatments does so by repeating or recapitulating the deliberate process of choice that produced these guesses in Phase I. There is, however, the possibility that some subjects simply have good memories; they may remember the numeric values of a large fraction of their guesses even if those guesses were not deliberate or systematic choices. In the *Memory* treatment, we provide a benchmark for how readily subjects can remember 20 guesses that do not follow any consistent system.

In Phase I of the *Memory* treatment, participants play 20 games against a computer that makes uniform random guesses. In each game, a subject was shown the computer's guess while making her own guess. The task was otherwise the same as in Phase I of the main treatments.[9] Phase II of the *Memory* treatment was identical to Phase II of the *Replicate* treatment; subjects were tasked with replicating their Phase I guesses but were not presented with the values of their Phase I guesses when prompted for their Phase II guesses. The number of remembered guesses in Phase II provides a benchmark for numeric memory.

2.3 Experimental Procedures

The experiment was run at Stanford with six, eight, or ten participants, all Stanford undergraduates, in each session. A session lasted about two hours, and subjects earned an average of \$55.17 including a \$5.00 show-up fee.

While subjects were initially informed of the two-phase structure, they received no details

[9] Phase I of the *Memory* treatment serves as an additional control, like Phase II of the *ShowGuesses* treatment, for determining whether participants are able and willing to calculate the best response to a known guess.

about Phase II until after Phase I was completed. After hearing instructions for Phase I, subjects completed an understanding test on paper followed by a second understanding test on the computer. Participants were given simple pocket calculators for use during the experiment.

For the first several decisions in each phase, subjects were not permitted to submit their guesses until after a certain time had elapsed; these restrictions were imposed in the hope that subjects would take the time to make thoughtful decisions.[10] After Phase II, subjects completed a short questionnaire and learned their monetary earnings from the experiment.

For each guess in a game, a subject could earn anywhere from 0 to 300 points. The point payoff function used was identical to that of CGC; it is a piecewise-linear decreasing function. Let $e_i = |x_i - y_i|$ denote the distance between participant i's guess x_i and her goal y_i in a certain game.[11] Then the points participant i earned from that trial were $s(e_i)$, where

$$s(e_i) = \begin{cases} 300 - \frac{11}{10}e_i & \text{if } e_i \leq 200 \\ 100 - \frac{1}{10}e_i & \text{if } 200 \leq e_i \leq 1000 \\ 0 & \text{if } e_i \geq 1000 \end{cases}$$

In the hope of mitigating concerns about unobserved varying risk preferences, the point payoffs in each trial were converted to realized money earnings using separate and independent binary lotteries run at the end of the experiment (Roth and Malouf, 1979). If a subject earned s points in a trial, the corresponding lottery paid \$2 with probability $s/300$ and \$0 with probability $1 - s/300$.[12]

[10] In Phase I, subjects had to wait 2 minutes for the first three trials and one minute for the next two trials before submitting a guess. In Phase II we employed similar timing restrictions: subjects had to wait one minute in each of the first five trials. For practical reasons (the experiment could not proceed to Phase II until all participants had completed Phase I), we also placed soft limits on the maximum amount of time subjects had to make decisions. In Phase I, this limit was five minutes for each of the first three trials, three minutes for each of the next two trials, and two minutes for each trial thereafter. In Phase II, subjects had up to three minutes for each of the first five trials and two minutes for each of the remaining fifteen trials. When the experimenter's screen showed a subject taking more than the maximum time, the experimenter made a verbal announcement reminding subjects to try to stay within the time limits. Otherwise, subjects could proceed at their own pace.

[11] In Phase I of the *BestRespond*, *Replicate*, and *ShowGuesses* treatments, $y_i = t_i x_j$, where x_j is the guess of the opponent and t_i is player i's target. For Phase I of the *Memory* treatment, x_j is the computer-generated guess shown to the subject while she chooses her own guess x_i. For Phase II, suppose x_i^I is i's guess from a given game in Phase I. In Phase II of the *Replicate* and *Memory* treatments, $y_i = x_i^I$. In Phase II of the *BestRespond* and *ShowGuesses* treatment, $y_i = t_i^{II} x_i^I$, where t_i^{II} is i's target in Phase II of that game. Note that y_i may fall outside of the guessing range $[l_i, u_i]$.

[12] In Phase I of the *Memory* treatment and Phase II of the *ShowGuesses* treatment, the winning lottery amount was \$1 instead of \$2, since these tasks were quite simple.

2.4 Predicted Behavior in Two-Person Guessing Games

To describe the equilibrium and behavioral game theory model predictions, we introduce the function $R(l, u, x) \equiv \min\{\max\{l, x\}, u\}$ (read, "restrict x to $[l, u]$"). That is, $R(l, u, x)$ is equal to l when $x < l$, u when $x > u$, and x otherwise. We select game parameters such that equilibrium play has a unique prediction.

Observation 1. (CGC) *Let $\{[l_i, u_i], t_i; [l_j, u_j], t_j\}$ be a two-player guessing game. When $t_i t_j \neq 1$ and payoffs are strictly positive, the game has a unique equilibrium (x_i, x_j) in pure strategies:*

If $t_i t_j < 1$, $x_i = R(l_i, u_i, t_i l_j)$ and $x_j = R(l_j, u_j, t_j l_i)$.
If $t_i t_j > 1$, $x_i = R(l_i, u_i, t_i u_j)$ and $x_j = R(l_j, u_j, t_j u_i)$.

Since we consider behavior in complete information games and focus on deterministic rules, the leading behavioral game theory models to describe subjects next to equilibrium are level-k and dominance-k. We adopt the common definition that a level 0 ($L0$) player i picks x_i randomly and uniformly from her action set. A player of level $k + 1$ believes the opponent uses the level-k rule and best-responds to this belief.

Here, an $L1$ player who best-responds to a hypothesized opponent who uniform-randomly chooses over her allowed guesses plays the same as if she believes her opponent will play the midpoint of her guessing range with certainty (see CGC); given strictly positive payoffs, the unique best-response is $R(l_i, u_i, t_i(l_j + u_j)/2)$. This pins down the behavior of higher levels in the level-k hierarchy: A level-k player chooses the best-response to the level-$(k - 1)$ guess of her opponent.

We also consider the dominance-k model examined by CGC, where hypothesized opponents randomly select among strategies that survive a certain number of rounds of iterated deletion of dominated strategies. We denote by $D1$ an agent who assumes that her opponent uniform-randomly selects among undominated guesses.

In general, let Dk be an agent who performs k rounds of iterated deletion and assumes her opponent uniform-randomly chooses a guess among the remaining actions.

In Table 2 we summarize the predicted guesses of the behavioral types we focus on in this paper. The table simplifies notation by shortening $R(l_i, u_i, x)$ to $R_i(x)$. The numeric values for the guesses of behavioral game theory types of the 20 games are given in Table 1.

2.5 Game Design

Each participant played a common set of twenty games, each with a unique Nash equilibrium. The games were chosen to identify various behavioral types. For each game, guessing range

Table 2.—Formulas for Behavioral Game Theory Types' Guesses

Strategy	**Formula for Player i**
Level 1	$R_i(t_i[l_j + u_j]/2)$
Level 2	$R_i(t_iR_j(t_j[l_i + u_i]/2))$
Level 3	$R_i(t_iR_j(t_jR_i(t_i[l_j + u_j]/2)))$
Equilibrium	$R_i(t_il_j)$ if $t_it_j < 1$ and $R_i(t_iu_j)$ if $t_it_j > 1$
Dominance 1	$R_i(t_i[R_j(t_jl_i) + R_j(t_ju_i)]/2)$
Dominance 2	$R_i(t_i[\max\{R_j(t_jl_i), R_j(t_jR_i(t_il_j))\} + \min\{R_j(t_ju_i), R_j(t_jR_i(t_iu_j))\}]/2)$

endpoints l and u were multiples of 50 between 0 and 1000, inclusive. Targets t were positive multiples of 0.1 in $(0, 1) \cup (1, 2)$.

While we use all eight games from CGC, of which one is a symmetric game, our subjects only play two instead of all eight from both sides, these constitute game 3, 4 and game 5, 6, see Table 1. In addition, we use two games where each has a dominant strategy for one player, these are game 19 and 20. For the remaining eight games, we wanted each game to provide type separation between the most common behavioral types, namely $L1$, $L2$, $L3$, $L4$, and EQ, to clearly identify the play of a subject. Otherwise, we had no specific hypotheses about what games were more or less conducive to behavior that concords with a given model. To ensure that we did not inadvertently choose parameters that favor certain behavior, we generated the remaining eight games randomly, subject to the above restrictions on the parameters and the requirement that there was a distance of at least 30 units between the $L1$, $L2$, $L3$, $L4$, and EQ predictions. These 8 games are games 11-18 in Table 1. As a result, in 14 games (8 randomly-generated and 6 from CGC) we have reasonable type separation between $L1$, $L2$, $L3$, $L4$, and EQ, with the distance between those types' predicted guesses never less than 10.5 units.[13] Type separation is helpful when aiming to classify subjects as the given types on the basis of observed choices.

3 Behavioral Types in Two-Person Guessing Games

Before we analyze the behavior of all 150 participants in the *BestRespond*, *Replicate* and *ShowGuesses* treatments, we provide evidence that our participants understand the games and seem sufficiently motivated by the incentives in the experiment.

We have 20 participants in the *Memory* treatment who in Phase I responded to the displayed guess of the computer, and 10 participants in the *ShowGuesses* treatment who in

[13] While 70% of games in our experiment have type separation between $L1$, $L2$, $L3$, $L4$, and EQ, in CGC this is the case for 50% of the 16 games. Partly as a result, we have fewer classified subjects than CGC. For details on the comparison between our results and those of CGC, see the online appendix.

Phase II responded to their Phase I guess while it was shown to them. Of those 600 guesses, all but 5 are within 0.5 units of the best response. This demonstrates that our participants understood the games and were willing and able to calculate the best responses to given guesses.[14] Being able to best-respond to a shown guess seems a minimal requirement for subjects who are supposed to form beliefs about the opponent and best-respond to them, which corresponds to the literal interpretation of the behavioral types we consider.

We also examine whether subjects made dominated guesses, which cannot be rationalized as best-responses to beliefs.[15] If subjects chose actions uniform-randomly in all Phase I decisions, we would expect 7.40 dominated guesses on average. In fact, the mean number of dominated guesses is 2.35 (s.d. 2.68).[16] About one-third of the 150 subjects (44) have no dominated guesses, and about two-thirds (97) have two dominated guesses or fewer. Only 17 subjects have 6 or more dominated guesses, and 9 of whom have 8 or more.

3.1 Behavioral Types Identified by the Apparent Classification

In this first part of the paper we use a simple and straightforward method to identify participants who can be described by $L1$, $L2$, $L3$, EQ, $D1$, or $D2$ on the basis of their Phase I play. We classify a participant i as having apparent type m when at least 8 of their 20 guesses (40 percent) were within 0.5 units of m_i, the action i would take under rule m. A 0.5 unit window ensures that a behavioral type guess m_i that is rounded to the closest integer is still counted as being a guess of behavioral type m.[17] While it is possible that a subject is classified as more than one type, this does not happen in our data.

With those parameters we classify 30% of participants; the results are shown in Table 3. A large fraction of classified subjects are EQ (10%) and $L1$ (9.3%) types, with $L2$ (6.7%) making up much of the remainder. We have not a single subject identified as $L4$ or $D2$, but have some matching $L3$ (2) and $D1$ (4). Starting from Nash equilibrium only, adding a small set of behavioral types increases the set of classified subjects by 200 percent. Compared to CGC, we have relatively more equilibrium types and somewhat fewer $L1$ and $L2$ types, for a more detailed comparison see the online appendix.[18]

[14] Furthermore, these results were obtained under experimental incentives half those used in the main treatments: in these trials, the lottery payoff was only one dollar instead of two.

[15] In a guessing game $\{[l_1, u_1], t_1; [l_2, u_2], t_2\}$ a guess x_i of player i is dominated if $x_i < \min\{u_i, t_i l_j\}$ or $x_i > \max\{l_i, t_i u_j\}$. Players are able to make dominated guesses in 13 and 17 of the 20 games, for G1 and G2 subjects respectively.

[16] Subjects do not appear to "learn" substantially over the course of the games by this measure: the number of dominated guesses in the first 10 games is 1.13 (s.d. 1.42) compared to 1.22 (s.d. 1.56) in the last 10 games, a small and statistically not significant difference ($p > 0.3$).

[17] In addition, a unit window of 0.5 and a match rate of 40% makes our analysis comparable with that of CGC; see Appendix.

[18] In CGC, of all classified participants, 46.4% are $L1$, 27.8% are $L2$, 4.7% $L3$ and 20.9% are EQ. They have

Table 3.—Summary of Estimated Type Distributions in Phase I

	$L1$	$L2$	$L3$	EQ	$D1$	$D2$	Uncl
Apparent	14	10	2	15	4	0	105
% of Classified	31.1%	22.2%	4.4%	33.3%	8.8%	0%	

Classifications of the 150 subjects using the apparent type classification method.

Because we allow subjects to make only small mistakes when matching against the model predictions, it is quite unlikely that a subject having eight guesses or more coinciding with a given model has this happen out of chance. There is, however, one behavior that may make a subject spuriously appear to match one of our behavioral types, the Nash equilibrium type. Specifically, for G1 and G2 subjects, 15 and 10 out of 20 equilibrium guesses are on the guessing range boundary, respectively. For the other behavioral types, at most 5 of the 20 predicted guesses are on the boundary. Hence, a player who always plays one of the boundaries might wrongly be classified as matching the equilibrium type. In our sample we may have misclassified two boundary-type players as equilibrium-type.[19]

The relative distribution of behavioral types among classified subjects is quite stable, even when using classification thresholds other than 40%. Specifically, for each subject and each behavioral type, we count the number of games where that subject's decision matches the behavioral type's prediction. We then identify the (perhaps non-unique) behavioral type with the largest count and call this the subject's modal type.[20] For any $q \in \{1, ..., 20\}$ and any behavioral type, we can compute the number of subjects whose modal type corresponds to the given model and who match that type in q or more games. For detailed results see the online Appendix.

While many subjects have fewer than 8 modal-type guesses, the non-modal-type guesses generally do not correspond to any of our other behavioral models. That is, subjects rarely "switch" from one behavioral type to another. Only 9% of subjects have more than 3 guesses matching behavioral types that are not their modal type.[21]

no $D1$ or $D2$ types when using the apparent type method.

[19] Of the 15 subjects identified with the equilibrium type, two have all their equilibrium guesses on the boundary, and furthermore, have 15 and 20 of their guesses on the boundary, respectively. The subject with 15 boundary guesses is from the *BestRespond* treatment and the subject with 20 boundary guesses is from the *Replicate* treatment. The other 13 equilibrium-type subjects have at most 10 guesses on the boundary and never more than 5 guesses on the boundary that are not equilibrium guesses. In addition, the difference in frequencies of hitting the equilibrium guess on the boundary versus the interior is never more than 65%.

[20] Since the modal type is in general not unique, we count a subject that has n behavioral types $\{m_1, ..., m_n\}$ for her modal type as $1/n$ of an m_1 player, for $m_i \in \{L1, L2, L3, EQ, D1, D2\}$ for $1 \leq i \leq n$.

[21] Subjects with 10–12 modal-type guesses seem to have the most behavioral type guesses that differ from their modal behavioral type. However, such subjects would be classified as their modal type given the apparent

4 Identifying Predictable Players

Behavioral game theory allows us to describe players using simple and portable models such as level-k and dominance-k. While the equilibrium type alone allows us to classify 10% of subjects, adding the level-k and dominance-k behavioral models brings this to 30%. This leaves almost 70% of subjects not classified as behavioral game theory types.

In this section we propose a test whether subjects deliberately play according to a deterministic rule. Basically, the test is whether subjects are predictable, that is, whether they behave in Phase II as expected given their Phase I choices. If a subject classified as a behavioral type is indeed truly paying according to that type, we would expect her to score highly on our test, that is, make Phase II guesses that conform to those expected given her Phase I guesses. If the 70% of subjects not classified as behavioral types are to a large extent not using deliberate deterministic rules, we would expect them, as a group, to not score highly on our test. We therefore expect subjects identified as behavioral game theory types to make Phase II choices that are more in concordance with their Phase I choices than unclassified subjects do. The flip side of this reasoning is that we aim to determine the success of existing behavioral types in identifying subjects who are using deliberate rules. As such, our approach will allow us to assess the scope for additional behavioral game theory models.

To evaluate whether a subject deliberately uses a deterministic rule, we exploit the expected relationships between Phase I and Phase II choices in our two main treatments. In the *Replicate* treatment, we expect any subject who uses a well-defined deterministic rule to be able to replicate her past behavior. In the *BestRespond* treatment, we expect a subject who in addition exceeds a minimal level of strategic reasoning to be able to best-respond to her past behavior. In contrast, a subject whose behavior is best described by a rule of thumb and hence does not include considerations about the opponents' behavior may fail to best-respond to her former actions. This may be the case even if she may be able to replicate her former actions.

In the last part of this section, we show that predictive success in Phase II cannot be explained merely by numeric memory of Phase I guesses. This shows that subjects who are classified as behavioral types are not conforming more to actions in line with their former behavior simply because they have "superior" memories. Furthermore, it suggests that subjects who were particularly successful in Phase II but whose Phase I actions were poorly matched by our behavioral models represent omitted types, rather than subjects with particularly good memories and no deliberate Phase I play.

type method anyway, as the threshold for classification is to have at least 8 guesses of the same type. Only 20% (21/105) of subjects with fewer than 8 modal types have a total number of 8 or more behavioral type guesses. The Figure in the online appendix shows for each number of modal type guesses n and each subject with n modal type guesses, the number of behavioral type guesses of the subject.

In this section we assess whether a subject behaves in Phase II in concordance with her Phase I guess by using a “model-free” “guess-level” approach. Specifically, we assume that the expected relationship between Phase I and Phase II behavior will manifest itself in each game. This model-free approach allows us to analyze whether the Phase II guess conforms to the profit-maximizing choice given the Phase I guess in that game, while remaining ignorant of the rules underlying Phase I choices. It allows us to compare the behavior of subjects we do classify as behavioral game theory types to those we cannot classify to any existing behavioral game theory type.

An obvious limitation of this guess-level approach is that we will fail to declare subjects as predictable when they implement rules in slightly noisy ways, or if they change rules in Phase I and cannot recall in Phase II when such changes were made. In Section 5 we will focus instead on a “rule-level” or “type-level” analysis, replacing the assumptions of the guess-level analysis with an assumed structural model of choice. The drawback of the approach in Section 5 is that it can be built only upon a known and limited set of behavioral types.

4.1 Can Players Replicate their Past Actions?

4.1.1 Classified versus Unclassified Participants

In the *Replicate* treatment, we have Phase I and Phase II observations for 63 participants.[22] In Phase II, participants are paid as a function of how close their guesses are to the guesses they made in Phase I. We say that a Phase II guess “replicates” the Phase I guess if the Phase II guess is within 0.5 units of the subject’s Phase I guess in the same game.[23] We call a subject a “replicator” if in at least forty percent of games (8 out of 20), their Phase II guesses replicate their Phase I guesses. Only 31 of the 63 subjects (49%) meet this criterion.[24] This criterion suggests that only half the subjects may be thought of as consciously and deliberately using a deterministic system of choice that could potentially be uncovered.

Of the 63 subjects in the *Replicate* treatment 18 (roughly thirty percent) were classified as matching one of our given behavioral types in Phase I; 5 as $L1$, 4 as $L2$, 1 as $L3$, 6 as EQ, and 2 as $D1$.[25] We find that 72% of these classified participants are replicators. The proportion of participants classified as level-k or dominance-k types who are replicators (8 of 12) is not significantly different from that of participants classified as the equilibrium type (5

[22] Subject 31 had a computer malfunction and could not finish Phase II; her data is dropped from this analysis.

[23] That is, x_i^{II} replicates x_i^{I} if $|x_i^{II} - x_i^{I}| \leq 0.5$, where x_i^{I} and x_i^{II} are the Phase I and Phase II guesses of participant i in the same game.

[24] Note that if all a subject recollects is that she played an action that was not dominated, we expect her to replicate only one guess, which corresponds to the game that has a dominant strategy.

[25] Of the 6 EQ players, one may be a misclassified boundary player.

of 6; $p = 1$). All proportion tests in this paper are Fischer's exact tests.[26] All comparisons of proportions of participants in this paper show p-values of two-sided Fischer exact tests. Of the 45 unclassified subjects, 18 (40%) are replicators. While this is not zero, it is significantly lower than the fraction of classified subjects who are replicators ($p = 0.01$). This suggests, first, that being able to replicate one's actions is not a trivial task for subjects. Second, subjects classified as behavioral types are strictly superior at this task, suggesting that behavioral game theory models have some success in uncovering subjects who use deliberate rules.[27]

To assess the extent to which behavioral models identify a large fraction of participants who deliberately use a deterministic rule, note that of the 31 replicators, only 42% are classified in Phase I. Specifically, we have 18 players who matched one of our behavioral models fewer than eight times in Phase I but who replicate 8 or more of their guesses. The fact that they can precisely replicate many of their non-behavioral type guesses suggests that these subjects are not merely subjects who play known behavioral rules with noise. They simply seem to follow different rules, suggesting room for additional behavioral game theory models.

In the following we show that the result that subjects classified as behavioral types are better at replicating their behavior than other subjects is robust when we use several different continuous measures to evaluate their success in replicating guesses.

Classified subjects have 11.88 replicated guesses, significantly more than unclassified subjects, 7.22 ($p < 0.01$). All tests of equality of means in this paper are t-tests. Furthermore, classified participants replicate overall 58% of their guesses, while unclassified subjects only replicate 36% of their guesses. The difference in the replication rate mostly stems from Phase I guesses that are behavioral type guesses. For such Phase I guesses the replication rate is 73% for classified subjects and 59% for unclassified subjects.[28]

We next assess the difference between classified and unclassified subjects by considering how far subjects are from replicating their guesses. For each subject i, we average—over the 20 games—the miss distance $|x_i^{II} - x_i^I|$, where x_i^I and x_i^{II} are the Phase I and Phase II guesses in a given game. Classified participants have a mean miss distance of 49.13, which is significantly lower than 74.28, the mean miss distance of unclassified subjects ($p = 0.036$).

The difference between classified and unclassified subjects manifests itself also in Figure 3. We order subjects by their average miss distances. We then plot the cdf of both classified and unclassified participants. Figure 3 shows that the 21 subjects with the lowest miss distances (the lower third) comprise 44 percent of all classified and 29 percent of all unclassified

[26] Likewise, the 5 $L1$ subjects are as likely to be replicators (3 out of 5) than the 5 that are $L2$ or $L3$ subjects (4 out of 5), $p = 0.99$.

[27] Note that even the 12 classified subjects who are not of the equilibrium type have a higher fraction of replicators (8 of 12) than the 45 unclassified subjects (18 of 45), though the difference is not significant, $p = 0.12$.

[28] For non-behavioral-type guesses in Phase I, classified participants replicate 23% of guesses, compared to 29% for unclassified subjects.

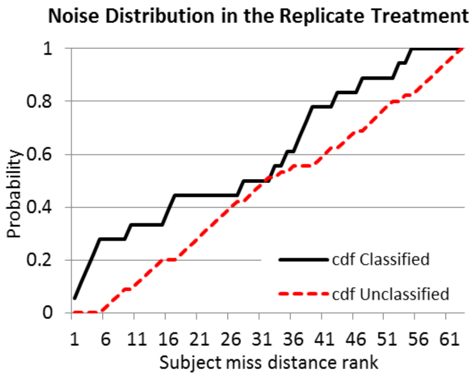

FIGURE 3.—The cdf of the 18 classified and the 45 unclassified participants in the *Replicate* treatment ordered by their miss distances. Subject 1 has the lowest miss distance, and subject 63 the highest.

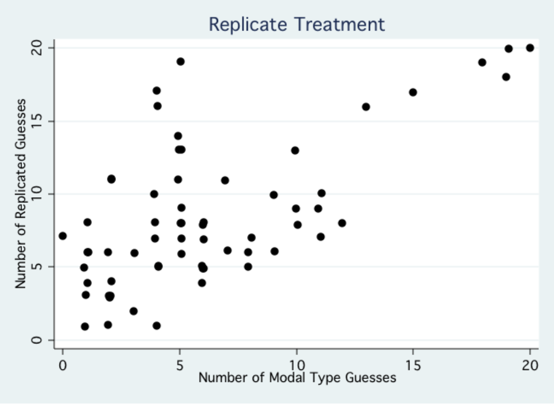

FIGURE 4.—For each subject, we plot the number of Phase II guesses that are replications of the corresponding Phase I guesses as a function of the subjects' number of modal type guesses in Phase I.

participants. The fact that the cdf of classified participants is above the cdf of unclassified participants reflects that classified participants have lower miss distances. The fact that the cdf of unclassified participants is not too far off the 45 degree line suggests that some unclassified participants are not much worse at replicating their choices than are classified participants.[29]

To compare the miss distances of subjects both within a treatment, but especially across treatments, we introduce a baseline miss distance. A subject who follows the "sophisticated rule" is a subject who in Phase II has no recollection of the guesses she made in Phase I, apart from the fact that the Phase I guess was not a strictly dominated guess. The sophisticated rule subject then randomizes in Phase II over any guess that is a best response to a surviving Phase I guess. In the *Replicate* treatment, the sophisticated rule therefore corresponds to randomizing in Phase II over guesses that are not strictly dominated. If subjects were to use the sophisticated rule, the mean miss distance of classified subjects would be 137.58, which is not significantly different from the mean miss distance unclassified subjects would have, 135.81 ($p = 0.846$). This suggests that the observed differences in miss distances between classified and unclassified subjects are not mechanically driven by the structure of the games or their different Phase I actions.

To assess the miss distance of a subject, we ask what fraction of reduction in miss distance a subject achieved compared to the sophisticated baseline. A subject who has the same miss distance as the sophisticated baseline has a reduction of 0, while 1 corresponds to a subject

[29] To provide some idea as to the miss distances, note that the lowest miss distance is 0, that of the 25th percentile is 36, the 50th is 61, the 75th is 91 and the highest miss distance is 211.

whose Phase II choices are payoff maximizing. Specifically, for each game i, let $MissDist_i$ be the distance between the subject's Phase II guess and the Phase I guess, and let $Soph_i$ be the (expected) distance between the Phase II guess and the Phase I guess under the sophisticated baseline rule. Then, we define $(Soph_i - MissDist_i)_+ \equiv \max\{Soph_i - MissDist_i, 0\}$ as the reduction in miss distance of the actual guess relative to the sophisticated baseline in game i. In a game in which $Soph_i > 0$, $(Soph_i - MissDist_i)_+/Soph_i$ is a value between 0 and 1 representing the normalized gains a subject made towards optimal play (a miss distance of zero) relative to the sophisticated baseline.[30] Losses are counted as zero gains.[31] For the set of games in which $Soph_i > 0$, we compute the mean of $(Soph_i - MissDist_i)_+/Soph_i$, yielding a measure of the gains towards optimal play in Phase II relative to the sophisticated baseline. On average, classified participants realize a significantly greater fraction of the gains towards optimal behavior than do unclassified participants, 71.6% versus 56.9% ($p = 0.009$).[32]

The fact that classified participants exhibit lower miss distances than unclassified participants is reflected in the earnings of subjects. Classified participants have 12% higher expected earnings than unclassified participants in Phase II, \$34.47 compared to \$30.71 ($p = 0.005$).[33]

4.1.2 Performance by number of modal type guesses

We show that the conclusions hold when, instead of partitioning subjects in Phase I into classified and unclassified sets, we consider how often subjects play their modal types, that is, their most-often played behavioral types. Figure 4 shows that the more modal type guesses a participant made in Phase I, the more guesses the participant replicates in Phase II. A regression of the number of replicated guesses in Phase II on the number of Phase I modal type guesses shows a coefficient of 0.670 (s.e. 0.101, $p < 0.001$) and a constant of 4.347 (s.e. 0.784, $p < 0.01$). The figure also shows that there are clearly many omitted types: subjects who

[30] $Soph_i$ is zero in the game with a dominant strategy.

[31] Note that when $Soph_i - MissDist_i$ is negative, dividing by $Soph_i$ does not normalize the losses to be between 0 and 1, and indeed they can take very high negative valuations, especially when $Soph_i$ happens to be small. Since that may distort the measure that considers average gains from the sophisticated baseline towards optimal play, we decided to count losses as zero.

[32] Furthermore, for each subject we can assess whether their miss distances are significantly different than the sophisticated baseline. Using a significance level of 10%, all classified subjects have significantly lower mean miss distances than the sophisticated baseline; this is the case for only 82% of unclassified subjects, a significant difference ($p = 0.092$). As we might expect, of the 31 subjects who are replicators (successfully replicated in 8 or more games), all have significantly lower miss distances, while this is the case for only 75% of the 32 non-replicators ($p = 0.004$).

[33] The maximum possible expected earnings in Phase II are \$40.00 for both groups of participants, which can be achieved for all possible Phase I actions. Note however that classified participants have significantly lower expected earnings from random uniform play than unclassified participants: \$14.90 and \$16.17, respectively ($p = 0.005$). Both higher average earnings and lower earnings from random play imply that classified participants realize a significantly larger fraction of the gains from optimal play relative to random play than unclassified participants, 78% compared to 60% ($p = 0.002$).

replicate their past guesses often while having few modal type guesses. That is, a sizable number of subjects seem to play according to rules they can replicate, while these rules do not match any of the behavioral models we consider. Precise replication of many guesses suggests that these are not subjects who implement a known behavioral type with noise.[34]

The conclusions are mirrored when we consider earnings. A regression of expected earnings in Phase II on the number of modal type guesses in Phase I shows a coefficient of 0.551 (s.e. 0.109, $p < 0.01$) and a constant of 28.33 (s.e. 0.846, $p < 0.01$). That is, each additional modal type guess in Phase I is associated with an increase in earnings of about 50 cents.

To summarize, we found that participants classified in Phase I by the method of Section 3 are to a large extent able to replicate their past guesses, confirming their behavioral type classification. Furthermore, participants who are not classified in Phase I successfully replicate in Phase II at a much lower rate, showing that existing behavioral models identify subjects who are more deliberate in their strategic choices. Finally, among participants who are replicators, 42% are classified, which suggests that quite a few participants who cannot be described by one of our behavioral types are nonetheless playing according to deterministic rules they can replicate. This suggests considerable room for new behavioral types. In this paper we call such rules, or types of players, omitted types.

4.2 Can Players Best-Respond to their Past Actions?

4.2.1 Classified versus Unclassified Participants

While being able to replicate a guess is consistent with the use of a deliberate rule, it does not necessarily indicate that a participant forms beliefs about the behavior of the opponent and then best-responds to those beliefs. Indeed, if the interpretation of the level-k type as an "as if" representation of a rule of thumb is accurate, we would expect level-k players to successfully replicate their past actions, but not necessarily best-respond to them. This would be also expected if the obtained level k is an indication of cognitive limitations. Furthermore, it remains an open question whether the omitted types we found in the previous section are best described by rules of thumb or by strategic rules that require subjects to be aware that they should first form beliefs about the behavior of the opponent and then best-respond to them. The goal of the *BestRespond* treatment is to shed light on these questions.

We have 76 participants in the *BestRespond* treatment, all of whom play the twenty games of Phase I in the same order once more but now in the role of the opponent. A subjects plays against a computer whose guess is her own past guess in that game. We call a Phase II guess a

[34] Recall that subjects with a low modal type have also few total behavioral type guesses. That is, these are also not subjects who simply switch among several known rules.

best response guess if it is within 0.5 units of the (unique) best response to the corresponding Phase I guess. That is, x_i^{II} is a best response guess if $|x_i^{II} - BR(x_i^I)| \leq 0.5$, where x_i^I and x_i^{II} are the Phase I and Phase II guesses of participant i in the same game $\{[l_1, u_1], t_1; [l_2, u_2], t_2\}$, and $BR(x_i^I) = t_2 x_i^I$ if $l_2 \leq t_2 x_i^I \leq u_2$, $BR(x_i^I) = l_2$ if $t_2 x_i^I \leq l_2$ and $BR(x_i^I) = u_2$ if $u_2 \leq t_2 x_i^I$. A participant is a "best-responder" if in at least forty percent of games her Phase II guess is a best-response guess. Only 31 of the 76 subjects meet this criterion. This suggests that only a small fraction of participants can be thought of as playing deliberate deterministic rules that are best responses to beliefs about the guess of the opponent.

In Phase I, roughly one-third (26 out of 76) of participants are classified through the method of Section 3; 9 are $L1$, 5 are $L2$, 1 is $L3$, 9 are EQ and 2 are $D1$.[35] We find that 81% of the 26 classified participants are best-responders. Level-k and dominance-k participants are as likely to be best-responders (13 of 17) as equilibrium participants (8 of 9) ($p = 0.614$). This suggests that the level-k model may be closer to an actual strategic description of behavior rather than merely an "as if" representation of a rule of thumb or cognitive limitation.[36]

Only 20% of unclassified subjects (10 of 50) are best-responders. While this is not zero, it is significantly smaller than the fraction of classified participants who are best-responders ($p < 0.001$). This shows that behavioral types are not only doing well in this strategic environment, but doing so is difficult.

To assess the extent to which behavioral models captured subjects who successfully best-responded, note that of the 31 best-responders in Phase II, 68% are participants classified as a behavioral type in Phase I. This suggests that existing behavioral models are particularly suited in identifying subjects who use deliberate rules that have some degree of strategic sophistication. We have, in addition, 10 participants who exactly best-respond to at least 40% of their guesses, but who were not classified as any behavioral type in Phase I. These subjects are prime candidates for omitted types, representing new behavioral models with a strategic element of agents best-responding to beliefs.

In the following discussion we show that the result that subjects classified as a behavioral type are significantly better at best-responding to their past behavior than others is robust when we use several more continuous measures of evaluating the success in best-responding.

When we compute the number of times participants best-respond to their past actions, classified participants on average have 11.42 best-responses, significantly more than the 5.46 of the unclassified participants ($p < 0.01$). Classified participants best-respond to 57% of all guesses compared to 27% of guesses for unclassified participants. This difference is mostly driven by the best-response rate to behavioral type guesses. For such Phase I guesses, the best-

[35] Of the 9 EQ players, one may be a misclassified boundary player.

[36] The 9 $L1$ subjects are somewhat less likely to be best-responders (5 out of 9) than the 6 that are $L2$ or $L3$ subjects (6 out of 6), though the difference fails to be significant, $p = 0.103$.

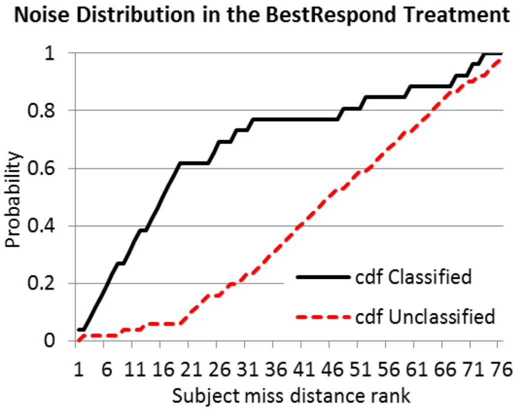

FIGURE 5.—The cdf of the 26 classified and the 50 unclassified participants in the *BestRespond* treatment ordered by their miss distances. Subject 1 has the lowest miss distance, and subject 76 the highest.

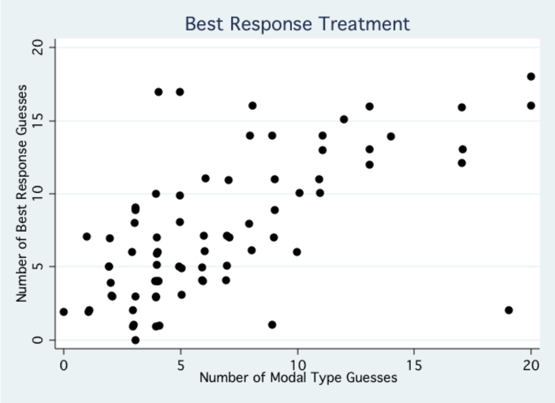

FIGURE 6.—For each subject, we plot the number of Phase II guesses that are best responses to the corresponding Phase I guesses as a function of the subjects' number of modal type guesses in Phase I.

response rate is 68% for classified participants compared to 35% for unclassified subjects.[37]

Alternatively, we can assess how well subjects best-respond to their past behavior by measuring how far away their guess in Phase II, x_i^{II}, is from the best-response to their Phase I guess, $BR(x_i^I)$. For each subject i we average over the twenty games the miss distance $|x_i^{II} - BR(x_i^i)|$. Classified subjects have a mean miss distance of 53.54, which is significantly smaller than the 97.90 miss distance of unclassified subjects ($p = 0.001$).

In Figure 5, we order subjects by their average miss distances. We then plot the cdf of both classified and unclassified participants. Figure 5 shows that the 25 subjects with the lowest miss distances (the lower third) comprise 65 percent of all classified and 16 percent of all unclassified participants. The fact that the cdf of classified participants is well above the 45 degree line, while that of unclassified participants is well below confirms that classified participants on average have lower miss distances, that is deviate much less from the best response than unclassified participants.[38]

In order to evaluate the miss distance and confirm that differences between classified and unclassified participants are not mechanically driven by their different Phase I play, we compute, as in the previous section, a baseline miss distance. A subject that follows the "so-

[37] For non-behavioral type guesses, the best-response rate is 33% for classified and 25% for unclassified participants. Subject fixed-effects conditional logit regressions confirm that guesses that are classified as best-response guesses are more likely to be best-responded to than are other guesses, and that this effect is significantly stronger for participants classified as a behavioral type.

[38] To provide some idea as to the miss distances, note that the lowest miss distance is 3, that of the 25th percentile is 35, the 50th is 74, the 75th is 117 and the highest average miss distance is 325.

phisticated" rule uniform randomly chooses among guesses that are best responses to Phase I guesses, for Phase I guesses that are not strictly dominated. The sophisticated baseline for classified subjects yields a mean miss distance of 105.03, which is not much lower than the mean miss distance for unclassified subjects using the sophisticated baseline 114.01 ($p = 0.111$). That is, the difference in the actual mean miss distances for these groups does not seem to be mechanically driven by differences in the structures of games or their Phase I play.

Finally, we assess what fraction of the reduction in miss distance from the sophisticated baseline to the optimum (zero miss distance) players achieved. As before, we normalize losses to 0, so that for each game the realized gains are normalized between 0 and 1. We take the average over all games in which the sophisticated baseline yields a strictly positive miss distance and average these measures separately over classified and unclassified subjects. Classified participants realize 68.7% of the gains towards optimal performance in Phase II relative to the sophisticated baseline, while this value is only 44.9% for unclassified participants ($p < 0.001$).[39]

The fact that classified participants tend to be closer to the Phase II best-responses than unclassified participants is also reflected in their Phase II expected earnings. Classified participants have 19% greater expected earnings than participants who are not classified ($30.13 compared to $25.36, $p < 0.001$).[40]

4.2.2 Performance by number of modal type guesses

We show that the conclusions hold when instead of partitioning participants on whether they had eight or more guesses of a behavioral type in Phase I, we use the number of guesses of their Phase I modal type. Figure 6 shows that the more modal type guesses a participant made, the more best responses the subject had in Phase II. A regression of the number of guesses best-responded to in Phase II on the number of modal type guesses of a subject in Phase I yields a slope coefficient of 0.634 (s.e. 0.092, $p < 0.01$) and a constant of 3.201 (s.e. 0.756, $p < 0.01$). Figure 6 also shows quite impressively the existence of subjects who are very good at best-responding to their Phase I guesses but who have few modal type guesses. These

[39] Furthermore, for each subject we can assess whether her miss distance is significantly different (smaller) at the 10% level than the sophisticated baseline. Of the 26 participants whose Phase I behavior classified them as behavioral types, 77% have significantly less noise in Phase II than had they used the sophisticated rule in Phase II. This is a significantly higher percentage than the 46% of the 50 unclassified participants ($p = 0.014$). As expected, of the 31 subjects who were classified as best-responders, 87% have significantly smaller mean miss distances than the sophisticated baseline, compared to only 36% of the 45 non-best-responders ($p < 0.001$).

[40] While the maximum possible expected earnings in Phase II are $40.00 for both groups of participants, this cannot be achieved for all possible Phase I actions, and differences in Phase I behavior across groups could mechanically produce differences in Phase II earnings. This does not seem to account for the difference we observe. The highest possible expected earnings are $36.14 and $35.94 for classified and unclassified participants, respectively ($p = 0.747$), while those for random play are $18.44 and $18.36 ($p = 0.828$). Classified participants realize 66% of the difference between random play and highest possible earnings, compared to only 38% for unclassified participants ($p < 0.001$).

subjects appear to represent omitted types. The fact that they so precisely best-respond to their guesses suggests that they play indeed omitted rules rather than being subjects who implement existing behavioral types with noise.

The conclusions are mirrored when we look at earnings instead of the number of best-response guesses. A regression of expected earnings on the number of modal type guesses a participant made yields a slope coefficient of 0.501 (s.e. 0.123, $p < 0.01$) and a constant of 23.60 (s.e. 1.015, $p < 0.01$). That is, each additional Phase I modal type guess is associated with a 50 cent increase in Phase II expected earnings.

We found that participants classified as behavioral game theory types in Phase I are to a large extent able to best-respond to their own past behavior, confirming their behavioral type classification. This also suggests that the interpretation of behavioral strategies as strategic choices might be more accurate than the interpretation that such models are largely “as if” models or rules of thumb. Furthermore, participants not classified in Phase I generally fail to best-respond to their past actions. That is, subjects who match behavioral types are clearly distinguished from those who do not. Finally, among participants who are best-responders, 68% are captured as behavioral types by the classification of Section 3. That is, behavioral strategies capture the majority of subjects who we judge as deliberate and strategic in this setting. There are, however, some omitted types, which suggests there is some room for additional strategic behavioral models.

4.3 Are More Participants Systematic Than Strategic?

The goal of this section is to assess to what extent subjects are more successful in replicating than best-responding to their past guesses. There are two reasons why replicating a guess may be easier than best-responding to it. While replicating a guess is clearly a necessary first step for best-responding to it, the latter entails, in addition, that the subject be aware that her action in a game should depend on her beliefs about the action of the opponent. A participant who uses a rule of thumb may never actually think about the opponent. She may not value the following information: the other player in Phase II of the *BestRespond* treatment is a computer who makes the subjects’ Phase I guesses. Therefore, subjects who use rules of thumb may be able to replicate their guesses but fail to make the strategic leap necessary to best-respond to them. A more mundane reason why best-responding is harder than replicating is that subjects now have an additional opportunity to make computational errors; once they compute the replication, they additionally must calculate the best response. Note, however, that we found that subjects make virtually no mistakes when computing the best responses to guesses. As noted in Section 3, out of 600 times subjects responded to shown guesses, all but 6 guesses were within 0.5 units of the best response.

4.3.1 Classified Participants

We first focus on classified participants, that is, participants who have guesses of the same behavioral type in 40% or more of the games. We saw that 72% of classified subjects were replicators and 81% were best-responders. This difference is not significant ($p = 1$). The number of Phase II guesses that correspond to the predicted guesses given the Phase I behavior is similar across treatments, it is 11.42 for subjects in the *BestRespond* treatment and 11.56 for subjects in the *Replicate* treatment ($p = 0.928$).[41]

For a continuous measure we report the mean miss distances of classified participants in the *Replicate* and the *BestRespond* treatments in Table 4 below. Classified participants have about the same mean miss distances in both treatments. Furthermore, they realize about the same gains towards optimal play relative to the sophisticated baseline (see (Soph-Miss Dist.)/Soph).

Therefore, for classified participants, best-responding to Phase I guesses seems no more difficult than replicating them. This suggests that not only the equilibrium type, but also the much more prevalent level k types are probably best thought of as strategic types rather than rules of thumb.[42]

	Replicate	*BestRespond*	t-test
Classified Subjects (N)	18	26	
Miss Distance	49.13	53.54	0.750
Sophisticated Rule	137.58	105.03	0.000
(Soph-Miss Dist.)/Soph	0.716	0.687	0.656
Unclassified Subjects (N)	45	50	
Miss Distance	74.28	97.90	0.021
Sophisticated Rule	135.81	114.01	0.000
(Soph-Miss Dist.)/Soph	0.569	0.449	0.003

TABLE 4.—Classified subjects: mean miss distances across treatments. The last column shows the p-values of two-sided t-tests of equal means across treatments.

4.3.2 Unclassified Participants

For subjects who are not classified in Phase I, we found that 39% are replicators, while only 20% are best-responders, a significant difference ($p = 0.046$). On average, unclassified subjects best-

[41] Across the two treatments, classified participants have about the same number of Phase I guesses that are classified, 13.88 for subjects in the *BestRespond* treatment and 13.83 for subjects in the *Replicate* treatment, $p = 0.964$.

[42] Even when we just concentrate on $L1$, or on all Lk types, such types are as likely to be best-responders as they are to be replicators across treatments: 5 of 9 and 3 of 5 ($p = 1$) for $L1$ and 11 of 15 and 7 of 10 ($p = 1$) for all Lk types.

respond to significantly fewer guesses than they replicate, 5.46 compared to 7.24 ($p = 0.028$). This difference is not driven by a difference in the number of behavioral type or modal-type guesses in Phase I across the *BestRespond* and *Replicate* treatments.[43]

Finally, we can compare the miss distances of subjects not classified in Phase I across treatments. In concordance with the results so far, unclassified participants in the *Replicate* treatment average significantly smaller miss distances than unclassified participants in the *BestRespond* treatment; the difference is almost 25%. Note that this is the reverse of the relationship between the sophisticated baseline measures, which suggests that this difference in actual miss distances is not mechanically driven by the structures of the games or tasks across treatments. Similarly, unclassified subjects in the *Replicate* treatment realize more of the gains towards optimal play relative to the sophisticated baseline compared to unclassified participants in the *BestRespond* treatment.

That for unclassified subjects replicating past behavior is so much easier than best-responding to it, and that there are fewer best-responders than replicators, suggest some of the omitted types may be better described by rules of thumb than by strategies that entail best responses to beliefs.

4.4 Playing according to a rule, or simply remembering guesses?

One interpretation of participants replicating and best-responding to guesses is that participants play according to some rule and are aware of it. However, it could be that some participants merely have good memories. In the *Memory* treatment, we provide a benchmark for how easy it is to remember 20 guesses that are not the result of any deliberate rules. To this end, in Phase I of the *Memory* treatment, participants play the guessing games against computers that make random guesses. A subject sees the opposing computer's guess while making her own guess. Like in Phase I of the other treatments, subjects are paid as a function of how far their guesses are from their goals. In Phase II of the *Memory* treatment, participants are tasked with replicating their Phase I guesses.

As in our other treatments, we use a 0.5 unit window to determine whether or not an individual guess is successfully "remembered" and say a subject is a "rememberer" if they have 8 or more successfully remembered guesses. In Phase II of the *Memory* treatment, subjects remembered between 1 and 7 guesses, so no subject can be said to be a "rememberer". On average, subjects remember 3.9 guesses out of 20. To assess that number, we compute the expected number of "remembered" guesses from a participant who has no memory, but plays

[43] In the *BestRespond* treatment, in Phase I, unclassified subjects have on average 5.26 behavioral type guesses and an average of 3.98 modal-type guesses, which is not significantly lower than the 4.89 ($p = 0.464$) and 3.71 ($p = 0.480$) in the *Replicate* treatment, respectively.

the action that would yield the highest expected earnings. The only available information for the participant is that in Phase I, the opponent—the computer—chose a guess uniform-randomly over the action set and the subject best-responded to that guess. This implies for each game a unique action in Phase II that maximizes expected earnings. If all 20 subjects would have used this scheme, they would have "remembered" on average 2.8 guesses. Note that while not much smaller, this is significantly different than the mean of 3.9 remembered guesses ($p = 0.03$).[44]

Finally, we can compare how well subjects perform in the *Memory* compared to the *Replicate* treatment. While 31 out of 63 subjects replicated 8 or more guesses and hence were replicators, no subject in the *Memory* treatment remembered 8 or more guesses, a significant difference ($p < 0.01$). The mean number of remembered guesses in the *Memory* treatment, 3.9, is also significantly lower than the average number of replicated guesses in the *Replicate* treatment, 8.5, ($p < 0.01$). Subjects in the *Memory* treatment even remembered fewer guesses than unclassified subjects were able to replicate in the *Replicate* treatment, 7.24, $p = 0.001$. A comparison to the *BestRespond* treatment yields similar results.[45]

That is, participants who perform well in the *Replicate*, or even in the *BestRespond* treatment, are unlikely to be subjects who merely have good memories. Rather, successful Phase II play likely involves reimplementation of deliberate Phase I rules.

5 Exploring the Boundaries of Classification

In the analysis so far, the criterion for whether a subject played according to a deliberate rule involved assessing whether a game's Phase II choice was correctly predicted by the guess from the corresponding Phase I game. This "model-free" "guess-level" approach had the advantage of providing expectations for Phase II actions even for subjects whose Phase I behavior did not match any of the given behavioral game theory models.

In this section we use a "type-level" approach. Specifically, we use all 20 guesses to assign a type to a subject not only in Phase I, but also in Phase II. We can then, for each treatment, assess concordance of the Phase II behavior with that implied by the assigned Phase I type.

[44] When considering the distribution of *# remembered guesses - # guesses remembered using the optimal no-memory scheme*, the mean is 1.1, standard deviation is 2.23, and minimum and maximum values are -4 and 5.

[45] In the *BestRespond* treatment, 31 out of 76 subjects were best-responders, a significantly greater proportion than the 0 rememberers out of 20 subjects ($p < 0.01$). The mean number of best-response guesses, 7.5, is also significantly greater than the mean of 3.9 remembered guesses from the *Memory* treatment ($p < 0.01$). Subjects in the *Memory* treatment were even worse at remembering guesses than unclassified subjects were at best-responding to them in the *BestRespond* treatment, as they averaged 5.46 best-response guesses compared to 3.9 remembered guesses out of 20 ($p = 0.07$).

This may increase the fraction of participants classified as behavioral types in Phase I whose guesses in Phase II are in accordance with the expected relationship between the phases.[46]

We use two classification methods to categorize participants in their Phase I and Phase II play. The first is the apparent type classification method from Section 3.1. One of its drawbacks is its equal treatment of all non-behavioral type guesses. In other words, a guess just outside the 0.5 window of the behavioral type prediction is not treated differently from a guess much further from the window. We could loosen classification criteria by either considering a lower threshold in the number of guesses needed to be classified, or by allowing for more error than 0.5 when assessing whether a guess is a behavioral type guess. However, this quickly leads to subjects being classified as several types.

The second classification method leads to a unique classification for each subject as it has a more sophisticated error structure than the apparent type method. This allows us to loosen classification criteria. Specifically, we use a maximum likelihood estimation to classify subjects. If a subject's behavior is not estimated to be significantly different from random play at the 5% level, we call the subject unclassified.

Note that our MLE(5%) classifies many more subjects compared to the apparent type classification. This allows us to assess the tradeoff between classifying additional subjects on the one hand, and on the other, successfully capturing subjects whose classifications in Phase I and Phase II are congruent with that of strategic subjects. Specifically, we can assess whether the additional subjects captured by a weaker criterion, MLE(5%), conform less well to their expected behavior than subjects classified by the apparent type method (or less lenient maximum likelihood estimations). Finally, we examine the likelihood that a subject behaves as predicted in Phase II as a function of the likelihood of the subjects' Phase I MLE classification.

To evaluate the behavior of subjects on the type level, we focus on the *BestRespond* treatment. The most mundane reason for this is that we have more subjects in that treatment. More importantly, we have seen in the previous section that participants classified as behavioral types using the apparent type method are able to best respond to their actions. On the other hand, subjects not classified by the apparent type method have been shown to have a much harder time best-responding to their guesses versus merely replicating them. If we view not only the equilibrium model, but also level-k and dominance-k, as models where a subject forms beliefs about the behavior of her opponent and then best-responds to those beliefs, then we may want to insist that subjects we classify as behavioral players conform to the expected

[46] For example, consider a hypothetical subject who made $L1$ guesses in even-numbered games in Phase I but deviated in odd-numbered games. Suppose this subject then made $L2$ guesses in the first 10 games of Phase II, but deviated in the last 10 games of the *BestRespond* treatment. In our previous "guess-level" analysis, the best-response rate would be 25%. In terms of a "type-level" analysis, the subject would be classified as $L1$ in Phase I and $L2$ in Phase II, and using the "type-level" approach could legitimately be viewed as being at least somewhat able to best-respond to her former behavior.

relationship in the more demanding strategic environment of the *BestRespond* treatment. In the online appendix, we show that the results are somewhat similar, if attenuated, in the *Replicate* treatment, and we explain why we expect this to be the case.

5.1 Type Transition in the *BestRespond* Treatment

To assess whether participants classified as a behavioral type in Phase I turn into the corresponding type in Phase II, we need to expand the set of types for Phase II. Specifically, we predict that a Lk player turns into $Lk+1$, for $k = 1, 2$, and 3, EQ remains EQ, and Dk turns into the best response to Dk (denoted by $BRDk$ for $k = 1, 2$). We therefore consider a set of Phase II types including $L4$, $BRD1$, and $BRD2$ in addition to the Phase I types.[47] We classify subjects independently by their behavior in Phase I, and then by their behavior in Phase II.

5.1.1 Type Transition Given the Apparent Type Classification

Table 5 shows the type transitions of the 76 participants in the *BestRespond* treatment using the apparent type classification.[48] Rows give the classification of the Phase I behavior and columns the classification of the Phase II behavior; the cell entries represent the number of subjects with that pair of Phase I and Phase II classifications. The entries in bold correspond to subjects who follow the prediction implied by best-response.

BR	Uncl.	L1	L2	L3	L4	EQ	BRD1	BRD2	D1	D2	Ph I
Uncl.	**44**	1	3	–	–	1	–	–	1	–	50
L1	1	1	**6**	–	–	1	–	–	–	–	9
L2	–	–	–	**5**	–	–	–	–	–	–	5
L3	–	–	–	–	**1**	–	–	–	–	–	1
EQ	3	–	–	–	–	**6**	–	–	–	–	9
D1	–	–	–	–	–	–	**2**	–	–	–	2
D2	–	–	–	–	–	–	–	–	–	–	–
Ph II	48	2	9	5	1	8	2	–	1	–	76

TABLE 5.—*BestRespond* type transition using the apparent type classification

For example, of the 9 subjects classified as $L1$ in Phase I, 6 are classified as $L2$, 1 is classified as EQ, 1 as $L1$ and 1 is not classified in Phase II. In total, if we assume that an unclassified participant is expected to remain unclassified, then 64/76 or 84% of subjects change types as

[47] We also keep the original types in Phase II, as a subject who has no recollection of her rule may turn into $L1$ or $D1$ depending on her beliefs about her use of dominated strategies.

[48] Recall that a subject is classified as a certain behavioral type if at least 40% of her guesses are within 0.5 of the exact guesses made by that type.

expected.[49] When we restrict attention to participants who are classified in Phase I, 20/26, that is 77% of subjects have the expected type transition.

One open question from the previous section concerns the 19% of classified participants who are not classified as best-responders using the guess-level prediction. It could be that a large fraction conforms to the expected behavior once we use the type-level classification. This is not the case. Of the 26 classified participants 21 are best-responders in Phase II, 20 conform with the expected type in Phase II and 18 fulfill both criteria.

Finally, we can turn the prediction exercise around; of the 28 subjects classified in Phase II, 22, that is 79%, are subjects who were classified in Phase I, and all but two of those changed classification according to the best-response model.

Overall, the type-level analysis confirms the conclusion that around eighty percent of subjects identified as behavioral types conform with their expected behavior in the *BestRespond* treatment when using the apparent type classification. That is, to a large extent, behavioral types capture subjects whose behavior is consistent with pure strategies that entail best-responding to beliefs.

5.2 Type Transition given by the Maximum Likelihood Estimation

The apparent type classification method quickly leads to multiple classifications when relaxed and arguably fails to capture subjects who implement behavioral rules with noise. We therefore consider, following CGC, a maximum likelihood method where we estimate each subject's type. We assume that each subject's behavior is determined, with error, by one of the behavioral types, or is completely random.[50]

Since the procedure forces all 150 subjects to be identified with some behavioral type in $\{L1, L2, L3, D1, D2, EQ\}$, we follow CGC and test whether the subject's maximum likelihood type gives a significantly better fit than the null model of uniform random play over the guessing range. For 14 of our subjects, we cannot reject the null of uniform random play at the 5% level; we term them unclassified.[51]

Table 6 shows the classification of subjects using the maximum likelihood method excluding subjects who fail the specification test using a 5% significance level (MLE(5%)). The relative frequencies of the behavioral types are not changed dramatically compared to the apparent type method with the exception of EQ which is less represented in MLE(5%). However, there

[49] Note that an alternative option for an unclassified subject would be to "turn into" $D1$ since most subjects do not use dominated strategies, or into $L1$ if the subject believes she played randomly in Phase I. However, only two subjects have such a classification combination for Phases I and II.

[50] Although there are slight differences in notation, our maximum likelihood estimation procedure is identical to the one in CGC; for details, see the Appendix.

[51] If we instead use a significance level of 10%, 9 subjects are unclassified.

Table 6.—Summary of Estimated Type Distributions in Phase I

	*L*1	*L*2	*L*3	*EQ*	*D*1	*D*2	Uncl
Apparent	14	10	2	15	4	0	105
% of Classified	31.1%	22.2%	4.4%	33.3%	8.8%	0%	
MLE (5%)	50	30	2	31	19	4	14
% of Classified	36.8%	22%	1.5%	22.8%	14%	2.9%	

BR	UNC	L1	L2	L3	L4	EQ	BRD1	BRD2	D1	D2	Ph I
UNC	**3**	2	1	1	–	–	–	–	–	–	7
L1	1	4	**12**	–	–	2	1	1	2	1	24
L2	1	2	2	**7**	–	1	1	–	3	1	18
L3	–	–	–	–	**1**	–	–	–	–	–	1
EQ	1	1	1	–	–	**14**	–	–	–	–	17
D1	–	2	–	–	2	–	**2**	–	1	1	8
D2	–	–	–	–	–	–	–	–	–	1	1
Ph II	6	11	16	8	3	17	4	1	6	4	76

Table 7.—*BestRespond* type transition using the MLE(5%) classification

are many more subjects classified using MLE(5%). This highlights the need to understand the potential trade-off between classifying more subjects and in turn having a potentially lower rate at capturing subjects with the expected type transitions.

5.2.1 Type Transition using MLE(5%)

In the *BestRespond* treatment, we classified 69 of the 76 participants using MLE(5%). Table 7 shows the transition of each subject where we classify behavior in Phase I and Phase II separately.

For example, of the 24 subjects classified as $L1$ in Phase I, half, 12, are classified as $L2$ in Phase II. One subject is unclassified in Phase II, two are classified as EQ, one each as $BRD1$, $BRD2$ and $D2$, 4 as $L1$ and 2 as $D1$. When we assume that an unclassified participant is expected to remain unclassified, then 39/76, 51% of subjects have the expected type transition. When we restrict attention to players that are classified in Phase I, then 36/69, that is 52% have the expected type transition.

Finally we can turn the exercise around and assess whether Phase I play is as expected given Phase II play. Of the 70 subjects classified in Phase II as behavioral types, 66 were classified in Phase I. However, only 51% had a type transition in concordance with the best response model.

5.3 The boundaries between classification and type consistent transitions

MLE(5%) classifies 69 participants in Phase I, which is more than the 26 classified by the apparent type classification. Using the apparent type method, 20 participants behave as expected, for a rate of 77%. For the maximum likelihood method, 36 behave as expected, for a rate of 52%, which is significantly lower ($p = 0.036$, Fisher's exact test). All the 26 participants classified by the apparent type method in Phase I retain the same classification using MLE(5%), which in addition classifies another 43 participants. Of those additional 43 participants only an additional 16 (37%) conform with prediction (significantly less than 77%, $p < 0.01$).

There are three groups from which these additional classified participants may be drawn. The first can be viewed as subjects who implement behavioral rules with error, let's call them BehError. These are subjects we hope to find using a classification method that allows for more error in the implementation of the behavioral rule.[52] The second group are subjects who behave very noisily, almost randomly, but we fail to identify as being random. The final group are subjects with omitted types. These are subjects who deliberately employ deterministic rules different from the behavioral models we considered. They may therefore be misclassified if they are classified as behavioral types by MLE(5%). Recall that in the previous section we found 10 participants who are prime candidates for being omitted types.[53]

We use two approaches to assess whether the rate of 37% of expected transitions of the 43 additional participants captured by MLE(5%) is high or low. First, we compare the rate of transitions that conform to expectations of all 43 additional classified subjects–who may be BehError subjects, random subjects and/or (misclassified) omitted types–to the subset of 10 subjects we have clearly identified as omitted types. Of the 10 subjects who are classified as best-responders but not as behavioral types using the apparent type classification, all are classified in Phase I using MLE(5%), 4 as $L1$, 3 as $L2$ and 3 as EQ. This suggests that MLE(5%) may have substantially over-fitted subjects as behavioral game theory types. Of those 10 subjects, 4 (40%) change classification as expected in Phase II using MLE(5%). Therefore, a 40% accurate type transition rate is far from proof that a large fraction of these additional classified subjects are those we want to capture, namely BehError types, that is

[52] BehError subjects will be missed using the apparent type classification whenever errors are likely to be larger than 0.5. Likewise, even if such subjects best-respond on a rule level, but implement both Phase I and Phase II rules with independent noise, such subjects may not be classified as best-responders; in each game, they are very likely to be more than 0.5 points away from exactly best-responding to their Phase I guess including the realized noise in that game.

[53] These subjects best responded to 40% or more of their guesses, while having strictly less than 40% of their Phase I guesses captured the same behavioral type. For example, two subjects have 4 or 5 guesses of the same behavioral type, respectively, while best-responding to 15 or more of their guesses. We argue that these 10 participants are of omitted types.

behavioral game theory types who implement their type with noise.[54]

Second, we highlight the tradeoff between classifying subjects as behavioral types and identifying subjects whose behavior in Phase II conforms to the behavior expected by the classification of the subjects' Phase I play. We therefore first force every subject to be classified as one of the types $\{L1, L2, L3, EQ, D1, D2\}$. Denote by τ_i^I the type in Phase I that MLE attributes to i, i.e., that has the highest likelihood for subject i. For each subject i, we then determine the type τ_i^{II} that corresponds to the best-response to their Phase I type τ_i^I. Instead of simply determining whether or not τ_i^{II} has the highest likelihood in Phase II, we compute the likelihood $l(\tau_i^{II})$ that subject i behaves in Phase II according to τ_i^{II}. To do this, we force the Phase II type of every subject to be among $\{L1, L2, L3, L4, EQ, D1, D2, BRD1, BRD2\}$. We therefore have, for every subject i, the likelihood $l(\tau_i^I)$ that she plays according to the type she is classified as in Phase I, and $l(\tau_i^{II})$, the likelihood that her Phase II type corresponds to best-response to her Phase I type. For an easier comparison, Figure 7 shows $(-ln(l(\tau_i^I)), -ln(l(\tau_i^{II})))$ for every subject. A high likelihood therefore corresponds to a low positive number, where the likelihood to play a type τ with certainty corresponds to a $-ln(l(\tau)) = 0$. To visualize the tradeoff between classifying participants above and beyond subjects classified by the apparent type method from Section 3, we distinguish, in Figure 7, between subjects who were classified by the apparent type method (C) and those that were unclassified (UC).[55]

The main result from Figure 7 is that the likelihood with which a subject is identified as playing the best response to her Phase I type is quite lower – and therefore the $-ln$ of that likelihood higher – when we move beyond the subjects who were already classified by the apparent type method. The average of $-ln(l(\tau_i^{II}))$, 59.6, for the 26 subjects classified by the apparent type method, is significantly lower than 106.5, the average for the 50 subjects not classified by the apparent type method, a significant difference using the non-parametric Mann Whitney test, $p < 0.01$.[56]

[54] Of the 4 omitted types classified as $L1$ in Phase I, one turns into an $L2$ player, one into $L3$, one into $D1$ and one into $D2$ in Phase II. Of the 3 omitted types classified as $L2$ in Phase I, one remains an $L2$, one turns into $L3$ and the third one is classified as $D1$ in Phase II given the maximum likelihood estimation. Of the 3 omitted types classified as EQ in Phase I, 2 remain EQ and one turns into $L1$.

[55] 25 of the 26 subjects classified by the apparent type method correspond to the 25 subjects who have the highest likelihood to be associated with a behavioral type in Phase I. The 26^{th} subject identified by the apparent type method is the 30^{th} individual when ordered by the likelihood to be associated with a behavioral type in Phase I per MLE. All subjects classified as behavioral types by the apparent type method retain their classifications under MLE.

[56] Note that the 7 subjects deemed unclassified under MLE(5%) have an average of 116.56, economically somewhat, and statistically significantly higher than the 104.86 of the subjects classified only under MLE(5%) but not the apparent type method ($p = 0.02$ using a Mann Whitney test). Note that this does not change the significant difference between subjects classified only under the apparent type method and those classified in addition under MLE(5%), ($p < 0.01$ using a Mann Whitney test). The 7 subjects deemed unclassified by MLE(5%) correspond to the last 3 subjects and the 65^{th}, 62^{nd}, 61^{st} and 53^{rd} of the 76 subjects, when considering the ordering induced by likelihood $-ln(l(\tau_i^I))$ of playing the type subjects are associated with in Phase I.

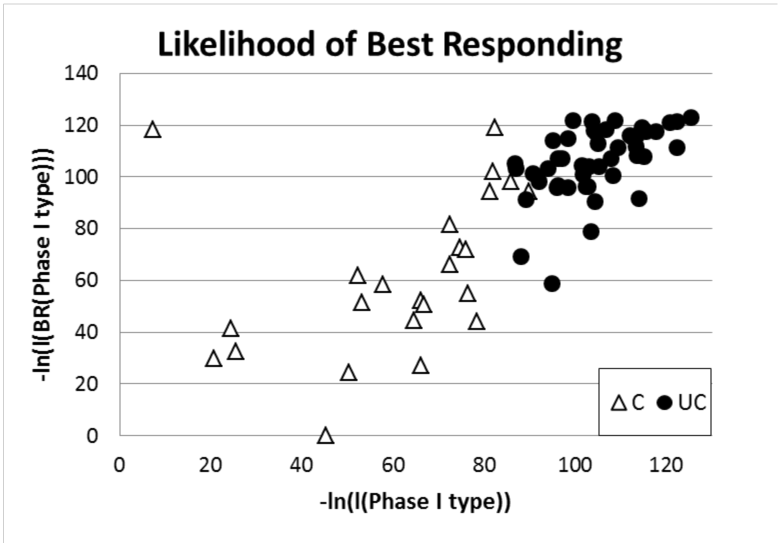

FIGURE 7.—For each subject i we plot the likelihoods $(-ln(l(\tau_i^I)), -ln(l(\tau_i^{II})))$ where τ_i^I is the type that corresponds to i's Phase I classification and τ_i^{II} is the best-response to τ_i^I. We distinguish between subjects who were classified by the apparent type method (C) and those that were not (UC).

We provide a second test to assess the extent to which subjects classified as behavioral types can be thought of as playing best-responses to the behavioral types associated with their Phase I play. Specifically, for each subject i, consider τ_i^{II}, the type that corresponds to the best-response to the classification of her Phase I behavior. We can compute the average distance over the 20 games between her guess in Phase II, x_i^{II}, and her expected guess given τ_i^{II}: $|x_i^{II} - \tau_i^{II}|$. We denote this as her type miss distance. Once more, while we use for every subject the Phase I type as estimated by MLE, we distinguish between subjects already classified by the apparent type method and others. This helps assess the tradeoff of classifying more subjects and capturing those whose behavior corresponds to best-responding to their Phase I types. The type miss distance of the 26 subjects classified by the apparent type method is 70.18, compared to 115.77 for the 50 other subjects, ($p < 0.01$).

Overall, the results point to the conclusion from before that as we move far beyond subjects already classified by the apparent type classification, subjects are less likely to exhibit behavior in concordance with their expected type in Phase II. this suggests that potentially only a small fraction of participants should be classified as behavioral game theory types. Those subjects implement their types with relatively little noise and are well described by their types even in additional strategic environments. While it is possible to fit additional subjects as behavioral game theory types given their Phase I behavior, such subjects are much less likely to behave as expected.

6 Related Literature, Methodological Remarks, and Discussion

The core approach of empirical game theory consists of observing strategic choices in specific settings. This has proven sufficiently powerful to topple several important null hypotheses, including the canonical model of unanimous Nash equilibrium play. However, these conventional strategic choice experiments offer limited power for delineating the set of subjects who play according to strategic models or determining the stability of behavior across strategic settings. In this section, we review previous methods and efforts to assess the stability of behavior and capture additional information on the processes generating strategic choices.

While the papers cited below have other valuable aspects we lack the space to discuss, we focus on the parts of papers that help determine the set of subjects who use deliberate rules and assist in understanding what sorts of rules these are; furthermore, we isolate aspects that concern the stability of behavior across settings, or, in other words, the degree of predictability of the behavior of subjects. We contrast these approaches with our design.

The most straightforward approach to assessing whether a subject identified as a certain behavioral type is “correctly” classified is to determine the stability of behavior out-of-sample. One possibility is to perform this exercise on the population level using different samples. This, however, may prove deceptively optimistic. For example, in our experiment, when we compare the distribution of types across Phases I and II, we would conclude—using the MLE(5%) classification in the *BestRespond* treatment—that 47 out of 69 in Phase I classified participants have associated partners in Phase II. Note that the number drops to 36 when requiring that the type transition be as predicted within participants, that is, the prediction rate drops from 68% to 52%.[57]

Most research on stability of rules within individuals has focused on comparing behavior across strategic settings.[58] Crawford and Iriberri (2007) look at various hide-and-seek games

[57] There are a couple of reasons why predicting play on the individual level is desirable. First, this may provide a more convincing test that the classification of a subject to a specific behavioral type is not erroneous. Second, we may aim to use individual characteristics such as demographics and intelligence measures to predict play. For approaches in this direction, see Burnham et al. (2009) for a positive correlation between depth of reasoning and IQ style measures, as well as Georganas et al. (2013) for a correlation of play with a CRT measure and Agranov, Caplin and Tergiman (2012) for a correlation between sophistication in the guessing game and a Monty Hall game. Another example is Coricelli and Nagel (2009), who correlate brain imaging results with depth of reasoning in a guessing game.

[58] An alternative method to assess type stability is to perform a hold-out prediction. This has been surprisingly unusual in the present literature with the exception of Stahl and Wilson (1995). They select a subset of games, estimate the subjects' type, and using the remaining games in addition, provide an estimate of the posterior probability that a subject has that particular type. When classifying a subject as stable if the posterior probability of having the same type is at least a (perhaps too modest) 15 percent, they find that 35 of 48 subjects are stable.

and find some consistency across games. On the other hand, Buchardi and Penczynski (2011) and Georganas, Healy, and Weber (2010) do not find strong consistency of play across guessing and hide-and-seek or "undercutting" games, respectively.

Failure to find type stability within a subject across strategic settings could be attributed to the subject being "erroneously" classified as a certain type. However, a lack of stability of a behavioral type can also be attributed to subjects having different beliefs about the behavior of others across different types of games. This poses inherent problems to out-of-sample predictions for models such as level-k of which one interpretation is that subjects best-respond to erroneous beliefs.[59] Indeed, our results from the *BestRespond* treatment suggest that level-k subjects are in general not rule of thumb players. There are several recent results that suggest that level-k subjects may not merely be rule of thumb players, Arad and Rubinstein (2012) and Agranov, Caplin and Tergiman (2012) (see also Georganas et al. (2013) below).[60]

To more precisely pin down rules underlying choice, researchers have worked to observe what parameters of a game are considered by subjects by hiding them and having subjects uncover each one individually (see Camerer et al. (1993), Costa-Gomes, Crawford, and Broseta (2001), Costa-Gomes and Crawford (2006), Brocas et al. (2010), and Wang et al. (2010). While this data can be very valuable and can rule out certain models of behavior, these approaches may not be inert with respect to the subjects' deliberations and could alter the strategic choice behavior we hope to observe.

Alternatively, researchers have tried to assess the thought processes with which decisions are reached through various communication devices. Most prominent is Burchardi and Penczynski (2012), where each of the two players in a team is randomly chosen to decide for the team. Before submitting choices, a subject can send a suggestion with explanations to her teammate. They find that roughly one third of subjects are non-strategic $L0$ players (see also Ball et al., 1991 and Sbriglia, 2008). Unfortunately, there is again a concern that the experimental paradigm may alter behavior.

Another approach has exploited the interpretation that behavioral models often rely on subjects holding erroneous beliefs about others, but that subjects otherwise behave in a profit

[59] Predictions would be more straightforward if those models were "as if" representations of rules of thumb.

[60] Arad and Rubinstein (2012) consider two versions of a game that only differ in the salience of $L0$ play. They find that while this manipulation does not increase the overall use of actions consistent with level k (for $k > 0$), it increased the frequency of actions associated with low levels of k. This is expected if the manipulation shifted not only the actual, but also the believed amount of $L0$ play. Agranov, Caplin and Tergiman (2012) observe choices in a version of the classic $\{[0, 100], 2/3\}$ guessing game. Subjects aim to guess 2/3 of the mean of 8 subjects who have already played the game. The innovation in that paper is to observe choices over the course of 3 minutes, where the decision at any second is potentially payoff relevant. They claim that about 57% of subjects are "strategic". Their choices average around 34 over the whole 3 minutes, but fall over time. Remarkably, they classify roughly 43% as naive - a fraction close to our findings. These subjects not only make average choices of 50 throughout the three minutes, their choices also do not fall over time.

maximizing way. This allows experimenters to assess those beliefs directly and check for payoff-maximizing behavior. Costa-Gomes and Weizsäcker (2008) show that elicited beliefs systematically conflict with their subjects' strategy choices; the beliefs suggest a greater strategic sophistication than the observed choices. In that vein, Bhatt and Camerer (2005) show differences in patterns of brain activation for corresponding belief elicitation and strategy choice tasks. One potential problem with this approach is that beliefs are in general elicited coincidentally with strategic choices, and as such may alter strategic thinking.[61] Alternatively, researchers have manipulated beliefs to determine whether the behavior of subjects changes accordingly. Georganas et al. (2013) manipulate subjects' beliefs about the strategic capacity of their opponent by providing information on their score on a battery of cognitive tests. They found that only some subjects adjust behavior in the expected direction. One possible explanation for the lack of change in behavior in the expected direction is that subjects—just like the authors—believe that the depth of reasoning of their opponent does not necessarily only depend on the cognitive abilities of the opponent, but rather on her beliefs about the degree of sophistication of others.

There is another paper, Ivanov, Levin, and Niederle (2010), that is initially similar to the present paper but reaches very different conclusions. Pairs of subjects bid in a common-value second-price auction. The experimenters first elicit the bid function in Phase I and observe, as expected, many subjects overbidding and facing the winner's curse, consistent with cursed equilibrium or a level-k model. Subjects then, in Phase II, face an additional set of auctions where the other player is replaced by an automaton that uses the subjects' Phase I bid function. They find that the Phase II bid function is not generally a best-response to the Phase I bid function. This is the case even though the subject gets to see her Phase I bid function while making her Phase II bids; that is, their experiment corresponds to our *ShowGuesses* treatment. It appears that in their common-value second-price auctions, subjects simply cannot (or are not willing to) compute best-responses to given bid functions. As such, their environment may be less amenable to models in which subjects hold erroneous beliefs about others, while behaving in payoff-maximizing ways given their beliefs. In our paper, we found that subjects are perfectly able to compute best-responses to given guesses; maintaining this assumption, our *Replicate* and *BestRespond* treatments then help elucidate the subjects' processes of strategic choice.

The part of our paper that assesses the stability of behavior is probably closest to out-of-sample prediction exercises. The main advantage of our approach is that if subjects are playing

[61] Several papers find that eliciting beliefs significantly alters play, see e.g. Rütstrom and Wilcox (2009), Erev, Bornstein, and Wallsten (1993), Croson (1999) and (2000), and Gächter and Renner (2010). Others fail to reject the null hypothesis that play is not affected by eliciting beliefs, e.g. Nyarko and Schotter (2002), and Costa-Gomes and Weizsäcker (2008).

according to a behavioral game theory type, we have precise expectations of their future play. A failure to comply with expected behavior in the *Replicate* treatment cannot be rationalized by, for example, subjects believing that as the number of games increases the opponent plays in a different way. Our two treatments are also uniquely suited to elucidate whether behavior that conforms with the level-k model (and dominance-k) is more likely an as if representation arising from a rule of thumb than an accurate description of participants strategically best-responding to non-equilibrium beliefs. Despite the precise test of whether subjects truly use a deterministic rule, we find very strong evidence and support not only for the equilibrium but also the level-k model.

7 Conclusion

To date there has not been a practical way to assess which players play according to a deliberate rule, especially one we have so far failed to identify, and which players do not. In this paper we say that a subject deliberately employs a well-defined rule if the behavior of the subject conforms to an expected relationship across strategic situations. We provide an environment and a test that allows us to identify such behavior. This allows us to relate existing behavioral game theory types with the set of subjects that use deterministic rules.

We augment choice data from a conventional strategic choice environment with information from treatments pitting subjects against their past behavior. We observe subjects' choices in two-player "guessing games"; we then surprise subjects by placing them in strategic situations where the optimal action depends on their previous choices. Subjects' play in the second phase of the experiment reveals the extent of their knowledge regarding how they arrived at their previously-made strategic choices. The design of our experiment allows us to provide a lower bound of how many subjects deliberately use deterministic rules.

The first environment where we assess this is the *Replicate* treatment, where we determine whether subjects can recreate their own actions in games. In a way, we assess whether subjects are predictable to themselves. Using specific thresholds to classify an action in a game as a replication and to identify subjects who are able to replicate their behavior, we found that roughly half the participants are replicators. The level-k model, jointly with equilibrium and the dominance-k model, account for one-third of subjects (of whom three-quarters are replicators). This suggests that there is noticeable room for additional behavioral models in accounting for subjects who are able to replicate their behavior.

In the *BestRespond* treatment, we require subjects to show strategic sophistication. We do this by paying subjects depending on how close they are to best-responding to their former actions. We find that there are much fewer subjects who are strategic than simply able to replicate their behavior. While only about 40% of subjects are best-responders, behavioral types comprise two-thirds of such subjects. Furthermore, behavioral types seem equally able to replicate and best-respond to their actions, while this is not the case for subjects not classified as behavioral types.

Overall, our results show that while equilibrium is able to account for two-ninths of strategic subjects, adding the level-k model brings this to almost two-thirds. We also have a small number of dominance-k subjects. Therefore, behavioral game theory has been quite successful in identifying strategic subjects. When considering only subjects who use well-defined deterministic rules they are able to replicate (rule-of-thumb players), there seems to be much more room for new behavioral models.

Lastly, this paper is also part of a small literature that tries to understand the "when" and "how" subjects think about opponents and contingencies (see Esponda and Vespa, forthcoming). We believe that our paper opens many avenues for future research. While we found type stability in our experiments, the stability of behavior across different types of games remains still unresolved. The results from our paper suggest that behavioral types are better interpreted as forming erroneous beliefs and best-responding to those beliefs than playing rules of thumb. As such, stability may only be found when assessing whether subjects are strategic per se.

8 References

Arad, Ayala and Ariel Rubinstein. 2012. "The 11-20 Money Request Game: A Level-k Reasoning Study"*American Economic Review*, 102 (7), 3561-3573.

Agranov, Maria, Andrew Caplin and Chloe Tergiman. "Nave Play and the Process of Choice in Guessing Games" `http://hss.caltech.edu/~magranov/documents/scp_revised.pdf`

Agranov, Maria, Elizabeth Potamites, Andrew Schotter and Chloe Tergiman. 2012. "Beliefs and Endogenous Cognitive Levels: an Experimental Study," *Games and Economic Behavior*, Vol. 75, pp. 449-463.

Ball, Sheryl B., Max H. Bazerman and John S. Carroll. 1991. "An evaluation of learning in the bilateral winner's curse," *Organizational Behavior and Human Decision Processes*, 48(1), 1-22.

Bernheim, B. Douglas. 1984."Rationalizable Strategic Behavior" *Econometrica* 52 (4), 1007–1028.

Blume, Andreas and Uri Gneezy. 2010. "Cognitive Forward Induction and Coordination without Common Knowledge: An Experimental Study," *Games and Economic Behavior*, 68, 2, 488–511.

Brocas, Isabelle, Juan D. Carrillo, Colin F. Camerer, and Stephanie W. Wang. 2010. “Imperfect Choice or Imperfect Attention? Understanding Strategic Thinking in Private Information Games.” `http://www.hss.caltech.edu/~sweiwang/papers/Betting.pdf`.

Burchardi, Konrad B., and Stefan P. Penczynski. 2012. “Out Of Your Mind: Eliciting Individual Reasoning in One Shot Games.” `http://people.su.se/~kburc/research/BurchardiPenczynski2012.pdf`

Burnham, Terence C., David Cesarini, Magnus Johannesson, Paul Lichtenstein, Bjrn Wallace. 2009. “Higher cognitive ability is associated with lower entries in a p-beauty contest,” *Journal of Economic Behavior and Organization*, 72, 171175.

Camerer, Colin F., Teck-Hua Ho, and Juin-Kuan Chong. 2004. “A Cognitive Hierarchy Model of Games.” *Quarterly Journal of Economics* 119 (3): 86198.

Cooper, David J., and John H. Kagel. 2005. ”Are Two Heads Better Than One? Team versus Individual Play in Signaling Games.” *American Economic Review*, 95(3): 477–509

Coricelli, Giorgio, and Rosemarie Nagel. 2009. “Neural Correlates of Depth of Strategic Reasoning in Medial Prefrontal Cortex. *Proceedings of the National Academy of Sciences*, 106 (23): 916368.

Costa-Gomes, Miguel A., Vincent P. Crawford, and Bruno Broseta. 2001. “Cognition and behavior in normal- form games: An experimental study.” *Econometrica*, 69(5):1193–1235.

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

Costa-Gomes, Miguel A. and Georg Weizsäcker. 2008. “Stated beliefs and play in normal-form games.” *The Review of Economic Studies*, 75(3):729–762.

Crawford, Vincent P., Uri Gneezy, and Yuval Rottenstreich. 2008.“The Power of Focal Points is Limited: Even Minute Payoff Asymmetry May Yield Large Coordination Failures” *American Economic Review*, 98(4): 1443–1458.

Crawford, Vincent P., and Nagore Iriberri. 2007. “Fatal Attraction: Salience, Naivete, and Sophistication in Experimental Hide-and-Seek Games.” *American Economic Review*, 97(5): 1731–1750.

Crawford, Vincent P., Costa-Gomes, Miguel A. and Nagore Iriberri. 2013. “Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications” *Journal of Economic Literature*, 51:1, 562.

Croson, Rachel T. A. 1999. “The Disjunction Effect and Reason-Based Choice in Games. *Organizational Behavior and Human Decision Processes*, Vol 80, 118-133.

Croson, Rachel T. A. 2000. “Thinking like a game theorist: factors affecting the frequency of equilibrium play", *Journal of Economic Behavior & Organization*, Vol. 41, 299314.

Erev, Ido, Gary Bornstein and Thomas S. Wallsten. 1993. "The negative effect of probability assessments on decision quality, *Organizational Behavior and Human Decision Processes*, 55, 78-94.

Esponda, Ignacio and Emanuel Vespa. forthcoming. "Hypothetical Thinking and Information Extraction in the Laboratory," *AEJ Microeconomics*.

Gächter, Simon, and Elke Renner. 2010. "The effects of (incentivized) belief elicitation in public good experiments," *Experimental Economics* 13(3), 364-377.

Georganas, Sotiris, Paul J. Healy, and Roberto A. Weber. 2013. "On the Persistence of Strategic Sophistication." `http://www.uni-bonn.de/~sgeorgan/mypapers/Manuscript.pdf`

Grosskopf, Brit and Rosemarie Nagel. 2008. "The Two-Person Beauty Contest", *Games and Economic Behavior*, 62(1), 93–99.

Ivanov, Asen, Dan Levin and Muriel Niederle. 2010. "Can Relaxation of Beliefs Rationalize the Winners Curse? An Experimental Study", *Econometrica*, Vol. 78, No 4, 1435-1452.

Nagel, Rosemarie. 1995. "Unraveling in Guessing Games: An Experimental Study," *American Economic Review*, 85, 13131326.

Nyarko, Yaw and Andrew Schotter. 2002. "1An Experimental Study of Belief Learning Using Elicited Beliefs," *Econometrica*, Vol. 70, No. 3, 971-1005.

Pearce, David G. 1984. "Rationalizable Strategic Behavior and the Problem of Perfection", *Econometrica*, 52(4), 1029–1050.

Rey-Biel, Pedro. 2009. "Equilibrium Play and Best Response to (Stated) Beliefs in Normal Form Games." *Games and Economic Behavior* 65 (2): 57285.

Roth, Alvin E. and Michael W. K. Malouf. 1979. "Game-Theoretic Models and the Role of Information in Bargaining", *Psychological Review*, 86(6), 574–594.

Rütstrom, Elisabet E. and Nathaniel T. Wilcox. 2009. "Stated beliefs versus inferred beliefs: A methodological inquiry and experimental test," *Games and Economic Behavior*, 67, 2, 616 - 632.

Sbriglia, Patrizia. 2008. "Revealing the depth of reasoning in p-beauty contest games," *Experimental Economics*, 11(2), 107-121.

Stahl, Dale O., and Paul R. Wilson. 1994. "Experimental Evidence on Players' Models of Other Players." *Journal of Economic Behavior and Organization*, 25 (3): 309–327.

Stahl, Dale O., and Paul R. Wilson. 1995. "On Players' Models of Other Players: Theory and Experimental Evidence." *Games and Economic Behavior*, 10(1): 218–254.

Wang, Joseph Tao-Yi, Michael Spezio, and Colin F. Camerer. 2010. "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation To Understand Truthtelling and Deception in Games." *American Economic Review*, 100(3): 984–1007.