Who is Strategic?*

Daniel E. Fragiadakis
Texas A&M University

Daniel T. Knoepfle
Uber

Muriel Niederle
Stanford University, *NBER* and *SIEPR*

December 11, 2016

Abstract

Behavioral game theory models are important in organizing experimental data of strategic decision making. However, are subjects classified as behavioral types more predictable in their choices than unclassified subjects? Alternatively, how many 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, seem to be the most prominent.[1] In a typical empirical paper, laboratory participants play a set of games

*We are especially grateful to Asen Ivanov for his impact on the design of the experiment. We thank Vince Crawford, Guillaume Fréchette, Matt Jackson and Emanuel Vespa 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 $k-1$.

and are then classified into a set of pre-specified behavioral types; see Crawford, Costa-Gomes, and Iriberri (2013) for an overview.[2]

Such a procedure, however, does not directly test whether the set of existing behavioral game theory types coincides with the set of participants who play according to a specific rule. For instance, a participant may follow a behavioral model that is yet to be discovered. In this paper, we develop a method to identify any participant who uses a specific rule, even if we do not know what that rule is. This allows us to determine how much room (if any) there is for new behavioral game theory models. Put differently, our test provides the fraction of subjects that additional models could conceivably capture. In this paper we focus on deterministic rules only. We discuss in the concluding section how our test might be expanded to include non-deterministic rules.

The set of participants classiffied as behavioral game theory types may differ from the set of participants who deliberately follow deterministic rules. First, due to the necessity of allowing for error when classifying subjects, a type I error can occur. That is, a participant can be misclassified as some behavioral game theory type when her underlying behavior follows a different rule or no rule at all. Second, a type II error can occur if a participant does not follow a pre-existing model but nonetheless deliberately applies a deterministic rule that we do not (yet) understand. In this paper we address both type I and type II errors. Specifically, we determine whether a subject belongs to one of the following sets (or neither or both): participants who follow existing game theory models and those who deliberately apply a deterministic rule. Using conventional methods, we can easily identify the first set. Determining who belongs to the second set is more challenging.

For example, if existing models fail to describe a participant's observed choices, we need to determine whether her decisions are reached via an unknown but otherwise deliberate process versus are chosen arbitrarily. To achieve this, we design a test to asses whether a participant deliberately uses a deterministic rule, be it a known rule from existing behavioral game theory models, or a rule for which no model yet exists. Furthermore, we test whether deterministic players are strategic. On the one hand, they may implement non-strategic "rules of thumb" that involve making actions in the absence of forming beliefs over opponent play. On the other, a deterministic player may follow a belief-based rule and be able to adapt her behavior to information about her opponent. Our approach provides insight into how much room there is for new behavioral game theory models and, in particular, the relative room for strategic belief-based models versus non-strategic rules of thumb.

In our experiment, subjects first play twenty two-player guessing games with anonymous

[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.

partners and without feedback (Phase I). The games are similar to those from Costa-Gomes and Crawford, 2006, henceforth CGC. Applying a conventional approach, we classify a participant as a behavioral type (from a set of models that includes equilibrium, level-k, and others) if it sufficiently explains a subject's gameplay. Under this approach, we classify 30% of subjects and leave the remainder unclassified. While this seems like a failure of our approach and/or existing models, it is unclear how many of the unclassified subjects we should expect to describe with yet to be developed models. In other words, if the unclassified subjects exhibit sufficient arbitrariness in their Phase I behavior, it would not only be extremely difficult to explain them with deterministic rules, doing so would simply be misguided. To confirm that these unclassified subjects are not merely confused, we have a control treatment, the *ShowGuesses* treatment, where we show that all subjects are virtually always able to best respond to any given opponent guess.

After completing Phase I, we present each participant with an unanticipated Phase II to test whether she deliberately applied a deterministic rule in Phase I. In Phase II of the *Replicate* treatment, a subject is tasked with replicating her Phase I behavior. Specifically, a participant is re-shown the Phase I games (with the order preserved) and is paid more as her Phase II guesses near her corresponding Phase I guesses. Under reasonable assumptions of self-awareness and cognitive ability, any subject who deliberately uses a well-defined deterministic rule in Phase I should be able to replicate it in Phase II, even if it is a rule of thumb. Conversely, arbitrary Phase I behavior should be non-replicable; indeed, results from a separate control treatment show that purely numeric memory of Phase I choices is very limited.

In Phase II of the *BestRespond* treatment, a subject is tasked with best-responding to her Phase I behavior. A participant replays the Phase I games (with the order preserved) but now takes the role that was previously occupied by her Phase I opponents. Furthermore, we inform a participant that her Phase II opponent is a computer that is programmed to make the Phase I guesses that she herself previously made; we do not inform the participant of her explicit Phase I choices. In effect, subjects in the *BestRespond* treatment play against their past selves.[3] Unlike the *Replicate* treatment, a participant's payoff-maximizing choice in a game in Phase II of the *BestRespond* treatment is the best-response in that game to the subject's original Phase I action. We reason that any subject who deliberately uses a belief-based rule in Phase I, and is aware of doing so, should be able to first replicate her former guess and then best-respond to it.

When considering the 70% of subjects that are unclassified in Phase I from the *Replicate* and *BestRespond* treatments, we find that most of them fail to meet a permissive threshold of

[3]This treatment is inspired by the design in Ivanov, Levin, and Niederle (2010) but has important differences we discuss below. A design in which subjects play against themselves is also a central component of Blume and Gneezy (2010).

making at least 8 optimal Phase II choices out of 20. The 30% of subjects that are classified, however, are far more likely to meet this threshold. This unequivocally confirms the success of existing behavioral models (such as level-k) in identifying the types of subjects that they intend to describe: participants that deliberately apply deterministic rules. Put another way, our test's results show that the classified subjects, as a group, are different from the unclassified subjects; these latter ones cannot be described as participants with equally well-defined decision rules. Furthermore, we find that classified subjects can best-respond to their former guesses just as well as they can replicate them while unclassified subjects find best-responding much harder than replicating. This result, coupled with the assumption that rule of thumb subjects should be able to replicate but not necessarily best respond to their behavior, suggests that non-strategic models would do better than belief-based rules in explaining the unclassified participants.

To further investigate this, we can consider subjects with high rates of replicating and best responding and ask how many are classified as behavioral types. In the *Replicate* treatment, existing models account for only 40% of subjects who score well in Phase II. In contrast, our existing models explain over two thirds of subjects who score well in Phase II of the *BestRespond* treatment. This difference provides further support to the hypothesis that there is more room for the development of new rules of thumb as opposed to novel belief-based models. In particular, 32% of subjects with high rates of best responding are unexplainable with our set of existing models while 35% are explained with level-k. Thus, a new class of strategic decision rules would at best describe a smaller proportion of subjects than level-k explains. Though we shed some light on the nature of new decision rules, our experiment was designed mainly to show their existence.

In summary, this paper provides a new methodology that can assess whether the behavior of an agent follows a deterministic rule versus exhibits idiosyncratic randomness in her decision-making. Importantly, the tests that we construct can identify deliberate subjects without having to understand the rules governing their choices. The value in capturing deliberate subjects before understanding their behavior is that it allows us to know how many (and which) subjects we should even attempt to describe with future models. The specific environment we consider in this paper involves pure strategies in two-player guessing games. In the concluding remarks, we discuss how our tests of deliberation might be expanded to mixed strategies as well as how they could be applied to not only games, but to non-strategic decision situations as well.

The paper proceeds as follows: Section 2 describes the experiment and Section 3 the classification of subjects. In Section 4 we present results pertaining to whether classified and non-classified subjects are able to replicate their previous actions and best respond to them. We

provide control treatments that show that the ability to replicate actions is due to recomputing a rule and not due to memorizing actual guesses. We also show that participants are able to compute the best response, hence failing to do so is not driven by a lack of understanding or computational ability. Section 5 discusses the literature and we conclude in Section 6.

2 The Experiment

2.1 Two-Person Guessing Games

Participants interact in simple complete information "two-person guessing games".[4] 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 strictly decreasing function $e_i = |x_i - t_i x_j|$, the realized distance from the player's guess x_i to her goal $t_i x_j$. 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.

2.2 Experimental Treatments

All treatments but one share a common two-phase structure. In Phase I, subjects play a series of 20 two-person guessing games against anonymous opponents without feedback. Game parameters are public information in all games and are 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.

	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

The experiment consisted of the *Replicate*, *BestRespond*, *ShowGuesses*, and *Memory* treatments. The Phase I tasks of the *Replicate*, *BestRespond*, and *ShowGuesses* treatments were

[4] 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 different models that have agents best-respond to beliefs result in different actions.

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 play all 20 games in individually-specific random orders without feedback on realized payoffs or opponents' guesses. Subjects are randomly and anonymously rematched with opponents before each game.[5] Subjects always see 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 is matched to opponent j in a given trial, she wishes to make a guess x_i as close as possible to her goal $t_i x_j$ and earns 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 faces the same sequence of 20 games from Phase I in the same order.[6] Participants are told that their goal in a Phase II game is to guess as close as possible to their previously made guess in that game in Phase I. In other words, for a given game, let x_i^I be the guess subject i makes in Phase I and x_i^{II} be her guess in Phase II. Then subject i's payoff from Phase II is strictly decreasing $e_i = |x_i^{II} - x_i^I|$ (according to the same function that translates her Phase I distance to her Phase I monetary payoff).

2.2.3 Phase II of the *BestRespond* Treatment

In Phase II of the *BestRespond* treatment, a subject faces the same sequence of 20 games from Phase I in exactly the same order. Now, however, subjects are informed they will play in the role of player 2, while the role of player 1 (that they had occupied in Phase I) would be taken by the computer. A subject is told that her computer will make the exact same guess that the subject previously made when playing the game in Phase I. Effectively, a participant plays against her former self. Mathematically, if subject i makes a guess of x_i^I in Phase I in game $\{[l_1, u_1], t_1; [l_2, u_2], t_2\}$, then her Phase II goal is to make a guess x_i^{II} that is as close as possible to $t_2 x_i^I$ (since t_2 is her target multiplier in Phase II). Subject i's payoff from Phase II is strictly decreasing $e_i = |x_i^{II} - t_2 x_i^I|$ (according to the same function that translates her Phase I distance

[5] Unknown to subjects, each participant plays either as Player 1 (P1) or as Player 2 (P2) in all Phase I decisions. See Table 1. Each pair of opposing players consists of one P1 subject and one P2 participant.

[6] The motivation behind the preservation of order across phases is 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 use each of their rules. Second, for every game in Phase I, the subject makes 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 pass 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	P1		100	900	0.5	150	250	112.5	100	162.5	131.25	1	
		P2		100	500	1.5	500	225	375	150	262.5	262.5		
		P1		300	900	0.7	350	546	318.5	300	451.5	423.15	2	
		P2		100	900	1.3	780	455	709.8	390	604.5	604.5		
		P1	P2	100	500	0.7	210	315	220.5	350	227.5	227.5	3	4
		P2	P1	100	500	1.5	450	315	472.5	500	337.5	341.25		
		P1	P2	300	500	0.7	350	420	367.5	500	420	420	5	6
		P2	P1	100	900	1.5	600	525	630	750	600	611.25		
	weak	P1		300	900	1.3	780	900	900	900	838.5	900	7	
		P2		300	900	1.3	780	900	900	900	838.5	900		
		P1		300	500	1.5	500	500	500	500	500	500	8	
		P2		300	900	1.3	520	650	650	650	617.5	650		
		P1		100	500	0.7	350	105	122.5	100	122.5	122.5	9	
		P2		100	900	0.5	150	175	100	100	150	100		
		P1		100	900	0.5	200	175	150	150	200	150	10	
		P2		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	P1		250	500	1.1	500	330	440	250	330	330	11	
		P2		150	950	0.8	300	400	264	200	300	276		
		P1		100	750	0.8	400	510	480	750	440	440	12	
		P2		50	950	1.5	637.5	600	765	950	637.5	652.5		
		P1		150	750	1.5	712.5	337.5	534.38	150	337.5	337.5	13	
		P2		50	900	0.5	225	356.25	168.75	75	225	178.12		
		P1		200	800	1.3	617.5	455	561.92	200	455	455	14	
		P2		50	900	0.7	350	432.25	318.5	140	350	324.8		
		P1	P2	250	1000	1.5	937.5	562.5	843.75	375	637.5	637.5	15	16
		P2	P1	250	1000	0.6	375	562.5	337.5	250	412.5	382.5		
		P1	P2	100	950	0.5	225	375	168.75	100	225	178.12	17	18
		P2	P1	150	750	1.5	750	337.5	562.5	150	356.25	356.25		
	weak	P1		350	500	0.5	350	350	350	350	350	350	19	
		P2		450	700	1.3	552.5	455	455	455	455	455		
		P1		450	850	1.1	467.5	660	660	660	676.5	676.5	20	
		P2		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. As an example, the targets of P1 and P2 in game 1 are 0.5 and 1.5, respectively. To conserve space, games with the same parameters (such as 3 and 4) are grouped together. For example, Player 1's lower bounds in games 5 and 6 are 300 and 100, respectively. 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.

to her Phase I monetary payoff). Subjects are not shown their previous Phase I guesses when making their Phase II guess. The games are presented to subjects as in Figure 2.

	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

To score well in Phase II of the *BestRespond* treatment, a subject has to understand that the information that her opponent is replaced by her computer that uses her Phase I choices is valuable. 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 subjects who succeed in replicating their Phase I guesses do not necessarily have choices in Phase I that are strategic, i.e. belief-based. 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.[7]

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

[7]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.

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 reimplementing 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. 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, a participant plays 20 games against a computer that makes a uniform random guess in each. A subject is shown the computer's guess before having to submit her own. Phase I was otherwise the same as in the other treatments.[8] Phase II of the *Memory* treatment is identical to Phase II of the *Replicate* treatment; subjects are tasked with replicating their Phase I guesses but are 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

Our study took place at Stanford University. Sessions consisted of either six, eight, or ten participants, all Stanford undergraduates. A session lasted about two hours, and subjects earned an average of $55.17 including a $5.00 show-up fee.

While subjects are initially informed of the two-phase structure, they receive no details about Phase II until after Phase I is completed. After hearing instructions for Phase I, subjects complete an understandings test on paper followed by a second computer-based understandings test. Participants are given simple pocket calculators for use during the experiment.

For the first several decisions in each phase, subjects are not permitted to submit their guesses until after a certain time elapses; these restrictions are imposed in hopes that subjects will take the time to make thoughtful decisions.[9] After Phase II, subjects complete a short

[8] 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.

[9] In Phase I, subjects have to wait 2 minutes for the first three games and one minute for the next two before submitting a guess. In Phase II we employ similar timing restrictions: subjects must 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 place soft limits on the maximum amount of time subjects can take to reach their decisions. In Phase I, this limit is five minutes for each of the first three trials, three minutes for each of the next two, and two minutes for each thereafter. In Phase II, subjects have up to three minutes for each of the first five trials and two minutes for each of the remaining fifteen. When the experimenter's screen shows a subject taking more than the maximum time, the experimenter makes a verbal announcement reminding subjects to try to stay within the time limits. Otherwise, subjects can proceed at their own pace.

questionnaire and learn their monetary earnings from the experiment.

For each guess in a game, a subject can earn anywhere from 0 to 300 points. The point payoff function used is 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.[10] Then the points participant i earns from that trial are $s(e_i)$, 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 hopes of mitigating concerns about unobserved varying risk preferences, the point payoffs in each trial are converted to realized monetary earnings using separate and independent binary lotteries run at the end of the experiment (Roth and Malouf, 1979). If a subject earns s points in a trial, the corresponding lottery pays \$2 with probability $s/300$ and \$0 with probability $1 - s/300$.[11]

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-

[10] 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. 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 $[l_i, u_i]$.

[11] 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.

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-$k+1$ player chooses the best-response to the level-k guess of her opponent.

We also consider the dominance-k model examined by CGC, where a Dk player performs k rounds of iterative deletion of dominated strategies and best responds the belief that her opponent plays uniformly at random among her 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.

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	$R_i(t_i R_j(t_j[l_i + u_i]/2))$
Level 3	$R_i(t_i R_j(t_j R_i(t_i[l_j + u_j]/2)))$
Equilibrium	$R_i(t_i l_j)$ if $t_i t_j < 1$ and $R_i(t_i u_j)$ if $t_i t_j > 1$
Dominance 1	$R_i(t_i[R_j(t_j l_i) + R_j(t_j u_i)]/2)$
Dominance 2	$R_i(t_i[\max\{R_j(t_j l_i), R_j(t_j R_i(t_i l_j))\} + \min\{R_j(t_j u_i), R_j(t_j R_i(t_i u_j))\}]/2)$

2.5 Game Design

Each participant plays a common set of twenty games, each with a unique Nash equilibrium. The games are chosen to identify various behavioral types. For each game, guessing range endpoints l and u are multiples of 50 between 0 and 1000, inclusive. Targets t are positive multiples of 0.1 in $(0, 1) \cup (1, 2)$.

We use all eight games from CGC, one of which one is a symmetric game. While CGC had subjects play all eight games from both sides, we did this for only two of the CGC games (see the 3-4 game pair and the 5-6 game pair in Table 1). In addition, we use two games (19 and 20) where each has a dominant strategy for one player. 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 a subject's. Otherwise, we had no specific

hypotheses about what games would be more or less conducive to behavior that concords with a given model. To ensure against inadvertently choosing parameters that favor certain behavior, we generate the remaining eight games randomly, subject to the above restrictions on the parameters and the requirement that there is 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.[12] Such type separation facilitates classifying a subject as a given type 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, are tasked with responding to the displayed guesses of the computer, and 10 participants in the *ShowGuesses* treatment who, in Phase II, are tasked with responding to their Phase I guesses while they are shown to them. Of those 600 guesses, all but 5 are within 0.5 units of the best response. This demonstrates that our participants understand the games and are willing and able to calculate the best responses to given guesses.[13] Being able to best-respond to a shown guess seems like 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 make dominated guesses, which cannot be rationalized as best-responses to beliefs.[14] If subjects chose actions uniformly at random in all Phase I decisions, they are expected to make 7.40 dominated guesses on average. In fact, the average number of dominated guesses is 2.35 (s.d. 2.68).[15] About one-third of the 150 subjects (44)

[12]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.

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

[14]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.

[15]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 ($p > 0.3$). We can analyze "learning" this way because every

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 in Phase I

We use a simple and straightforward method – very much in line with CGC – 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%) are 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. While it is possible that a subject is classified as more than one type, this does not happen in our data. While there is no theoretical reason for these two cut-offs, we use them as they mirror those of CGC. In the Appendix, Section 8.2, we show how the classification changes when we relax these cut-offs. Even though the number of classified participants varies, the relative distribution of types is quite stable.

With these 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. Not a single subject is identified as $D2$, but some match $L3$ (2) and $D1$ (4). We find no $L4$ subjects. 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.[16]

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

	$L1$	$L2$	$L3$	EQ	$D1$	$D2$	*Unclassified*
Organization of Subjects	14	10	2	15	4	0	105
Percentage of Classified	31.1%	22.2%	4.4%	33.3%	8.8%	0%	-

Subjects (150) are pooled from the *Replicate*, *ShowGuesses* and *BestRespond* treatments

Because we allow only for small mistakes when matching subjects 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 strategy (outside our pre-specified models) that may make a subject spuriously appear as a Nash equilibrium type. Specifically, for P1 and P2 subjects, 15 and 10 out of 20 equilibrium guesses are on the guessing range

subject sees the 20 games in a subject-specific random order.

[16]In CGC, of all classified participants, 46.4% are $L1$, 27.8% are $L2$, 4.7% $L3$ and 20.9% are EQ. They have no $D1$ or $D2$ types when using their "apparent from guesses" method.

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 there are two subjects for which this may be a concern.[17]

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.[18] For more details see the Appendix, Section 8.3.

4 Who uses a Deterministic Rule and Who is Strategic?

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. A traditional next step would be to relax classification criteria by allowing participants to implement their strategy with error. Such an exercise, in general, restricts attention to a given set of behavioral game theory types.[19] In contrast, a goal of this paper is to assess how many of the 70% unclassified subjects use deterministic rules that differ from those described by existing behavioral game theory models.

We therefore propose a test of whether or not a subject deliberately plays according to a deterministic rule. In short, the test checks whether subjects are predictable, that is, if they behave in Phase II as expected given their Phase I choices. If a subject who is classified as a behavioral type is indeed applying that type's rule deliberately, we would expect the participant to score highly on our test. This expectation is driven by the fact that, to have been classified in the first place, a participant must have implemented her behavioral type with essentially no

[17] Of the 15 subjects identified as equilibrium types, 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%.

[18] 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 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.

[19] There are, of course, exceptions, e.g. CGC).

error in at least 40% of the games. We therefore expect a classified subject to make Phase II guesses that conform to the predictions generated from her Phase I guesses (and the particular treatment she is in). 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 when compared to unclassified subjects. 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 such a subject who, in addition, exceeds a minimal level of strategic reasoning to be able to best-respond to her past behavior. A deliberate subject whose behavior is best described by a rule of thumb (that sidesteps considerations about the opponents' behavior) may be able to replicate but not best-respond to her former actions.

Our test relies on the assumption that participants recompute their guesses rather than remember their numeric values. 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 conforming more to actions in line with their former behavior because they are more likely to regenerate their guesses, presumably because they use deterministic rules, rather than because they have "superior" memories. Furthermore, it confirms that subjects who were particularly successful in Phase II but whose Phase I actions were poorly matched by our behavioral models are more likely to represent "omitted types" than arbitrary subjects with particularly good memories.

In this paper, we assess whether a subject behaves in Phase II in concordance with her Phase I guess by using a "guess-level" approach that is "model-free". 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 a Phase II guess conforms to the profit-maximizing choice given the corresponding Phase I guess while remaining ignorant of the rule underlying the Phase I choice. This lets us compare the cross-phase predictability of the group of classified subjects versus the unclassified participants.

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.[20] 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.[21] We call a subject a "replicator" if in at least 40% of games (8 out of 20), her Phase II guesses replicate her Phase I guesses. Only 31 of the 63 subjects (49%) meet this criterion.[22] This criterion suggests that only half the subjects may be thought of as consciously and deliberately using deterministic systems of choice that could potentially be uncovered.

Of the 63 subjects in the *Replicate* treatment, 18 (roughly thirty percent) are 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 $D1$.[23] We find that 72% of these classified participants are replicators. The proportion of classified 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 of 6; $p = 1$). All proportion tests in this paper show p-values of two-sided Fischer exact tests.[24] 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. 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.[25]

To assess the extent to which behavioral models identify participants who deliberately use deterministic rules, note that of the 31 replicators, only 42% are classified in Phase I. Specifically, we have 18 players who match each behavioral model fewer than eight times in Phase I but who replicate 8 or more of their guesses in Phase II. 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. Quite the contrary, they seem to be non-noisy followers of unknown rules, suggesting room for new behavioral game theory

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

[21] 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 respective Phase I and Phase II guesses of participant i in the same game.

[22] 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.

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

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

[25] 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, $p = 0.12$.

models to describe these "omitted types".

In the following paragraphs, we consider several different continuous measures of Phase II performance; we find robustness of our previous finding that classified subjects are better at replicating their behavior than unclassified subjects.

Subjects identified as behavioral types have significantly more replicated guesses compared to unclassified subjects: 11.88 versus 7.22 ($p < 0.01$). All tests of equality of means in this paper are t-tests. Furthermore, when considering all guesses, classified participants replicate 58% of them, while only 36% are replicated by unclassified subjects. 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.[26]

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 the mean miss distance of 74.28 held by the unclassified subjects ($p = 0.036$). This difference is also 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 ($p = 0.005$).[27]

The distinction between classified and unclassified subjects also manifests itself in Figure 3 which orders subjects by average miss distance and plots the cdfs 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 participants. The fact that the cdf of classified participants is above the cdf of unclassified participants reflects that classified participants have lower miss distances. 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 compared to classified participants.[28]

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

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

[27]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% ($p = 0.002$).

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

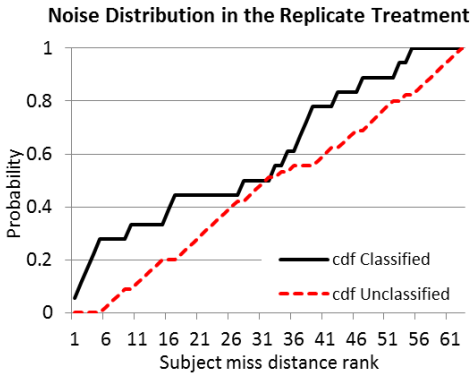

FIGURE 3.—The cdf of the 18 classified and the 45 unclassified participants in the *Replicate* treatment, ordered by their miss distances. Subject 1 has the lowest miss distance and subject 63 the highest.

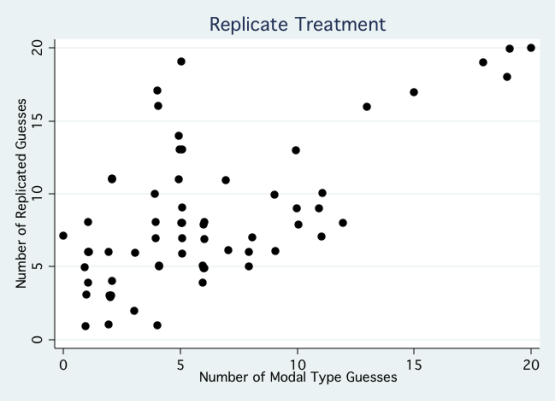

FIGURE 4.—For each subject, we plot the number of Phase II guesses that are replications of the corresponding Phase I guesses as a function of the number of modal type guesses in Phase I.

rule" is a subject who in Phase II has no recollection of the guesses she made in Phase I, apart from the fact that it 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 nor their different Phase I actions.[29]

[29]We can also ask what fraction of reduction in miss distance a subject achieved compared to the sophisticated baseline. We find that classified participants realize a significantly greater fraction of the gains towards optimal behavior than do unclassified participants. A subject who has the same miss distance as the sophisticated baseline has a reduction of 0, while 1 corresponds to a subject 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. $Soph_i$ is zero in the game with a dominant strategy. Losses are counted as zero gains. 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. 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

4.1.2 Performance by number of modal type guesses

In the following paragraphs we adopt a more continuous measure of Phase I behavior and show further robustness of our previous finding that classified subjects are better at replicating their behavior than unclassified subjects. Instead of partitioning subjects in Phase I into sets of classified and unclassified participants, we consider how often a subject plays her modal type, i.e., her most frequently played behavioral type. Figure 4 shows that the more modal type guesses a participant makes in Phase I, the more guesses she 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 often replicate their past guesses 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 noisily implement knowns behavioral types, nor are they subjects who simply switch among several known rules.[30]

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 find 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 classifications. Furthermore, as a group, participants who are not classified in Phase I successfully replicate in Phase II at much lower rates, showing that existing behavioral models identify subjects who are more deliberate in their 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.

---

average, classified participants realize a significantly greater fraction of the gains towards optimal behavior than do unclassified participants, 71.6% versus 56.9% ($p = 0.009$). 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 ($p = 0.004$).

[30] Recall that a subject with a low modal type also has few behavioral type guesses.

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 deliberate use of a deliberate deterministic 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. Most importantly, it remains an open question whether the unclassified replicators (omitted types) are best described by rules of thumb versus strategic rules that involve the formation of beliefs followed by best responses. The goal of the *BestRespond* treatment is to shed light on these questions.

We have 76 participants in the *BestRespond* treatment, who, in Phase II, play against their Phase I selves. Specifically, in Phase II, subjects play the 20 games of Phase I (in the same order), but take on the role of their Phase I opponent. A subject's Phase II opponent is her computer that plays in the participant's Phase I role and makes her exact Phase I guess. (The subject is informed that the computer is programmed this way, but is not explicitly shown her previous guesses.) We call a Phase II guess a "best response guess" if it is within 0.5 units of the unique best response to its corresponding Phase I guess. That is, x_i^{II} is a best response guess if and only if $|x_i^{II} - BR(x_i^I)| \leq 0.5$, where (i) x_i^I and x_i^{II} are the respective Phase I and Phase II guesses of participant i in the same game $\{[l_1, u_1], t_1; [l_2, u_2], t_2\}$, (ii) $BR(x_i^I) = t_2 x_i^I$ if $l_2 \leq t_2 x_i^I \leq u_2$, (iii) $BR(x_i^I) = l_2$ if $t_2 x_i^I < l_2$ and (iv) $BR(x_i^I) = u_2$ if $t_2 x_i^I > u_2$. 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 deliberately playing deterministic rules that are best responses to beliefs about the guesses of the opponents.

In Phase I, roughly one-third (26 out of 76) of participants are classified using the method of Section 3; 9 are $L1$, 5 are $L2$, 1 is $L3$, 9 are EQ and 2 are $D1$.[31] 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 are 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 as oppose to a mere "as if" representation of a rule of thumb or cognitive limitation.[32] Only 20% of unclassified subjects (10 of 50) are best-responders. While this is not zero,

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

[32] 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, $p = 0.103$.

it is significantly smaller than the fraction of classified participants who are best-responders ($p < 0.001$). These results show that behavioral types are not only doing well in this two-phase strategic environment; we see that performing well is difficult.

To assess the extent to which behavioral models capture subjects who successfully best-respond, note that of the 31 best-responders in Phase II, 68% are participants classified as behavioral types in Phase I. This suggests that existing behavioral models are particularly suited in identifying subjects who use deliberate rules that have some degrees of strategic sophistication. We have, in addition, 10 participants who provide exact best-responses to at least 40% of their guesses but who were not classified as any behavioral type in Phase I. These omitted types are prime candidates for being described with new belief-based behavioral game theory models.

In the following paragraphs, we consider several different continuous measures of Phase II performance; we find robustness of our previous finding that classified subjects are better at best responding to their behavior than unclassified subjects.

When we compute the number of times participants best-respond to their past actions, classified participants have, on average have 11.42 best-responses while unclassified participants have only 5.46; this difference is statistically significant ($p < 0.01$). Classified participants best-respond to 57% of all guesses compared to 27% for unclassified participants. This difference is mostly driven by the best-response rate to behavioral type guesses. For such Phase I guesses, the best-response rate is 68% for classified participants compared to 35% for unclassified subjects.[33]

Alternatively, we can assess how well a subject best-responds to her past behavior by measuring how far her Phase II guess (x_i^{II}) lies from the exact best response ($BR(x_i^I)$). For each subject i, we average this discrepancy ($|x_i^{II} - BR(x_i^i)|$) over the twenty games to compute i's average miss distance. The mean of classified subjects's average miss distances is 53.54; this is significantly smaller than the corresponding statistic of 97.90 unclassified subjects ($p = 0.001$). This difference is also reflected in the earnings of subjects: classified participants have 19% greater expected earnings than participants who are not classified ($30.13 compared to $25.36, $p < 0.001$).[34]

[33]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 behavioral types.

[34]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

In Figure 5, we order subjects by their average miss distances. We then plot the cdfs 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 mean miss distances; that is, classified subjects deviate much less from best responses than do unclassified participants.[35]

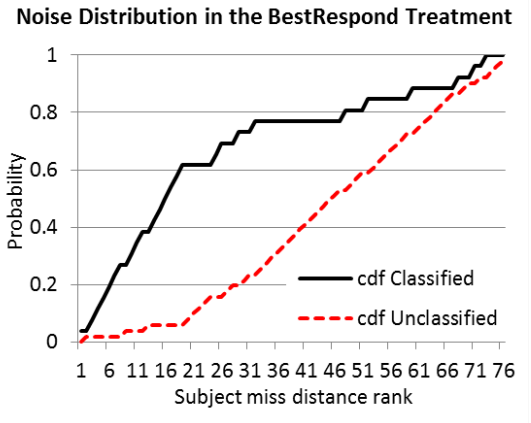

FIGURE 5.—The cdf of the 26 classified and the 50 unclassified participants in the *BestRespond* treatment ordered by their miss distances. Subject 1 has the lowest miss distance, and subject 76 the highest.

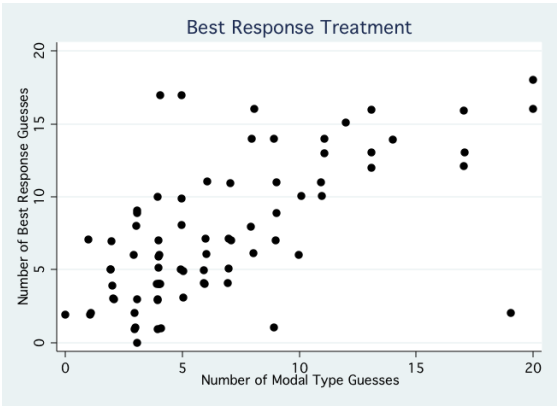

FIGURE 6.—For each subject, we plot the number of Phase II guesses that are best responses to the corresponding Phase I guesses as a function of the subjects' number of modal type guesses in Phase I.

In order to confirm that difference in average miss distances between classified and unclassified participants is not mechanically driven by these groups' different Phase I choices, we compute, as in the previous section, a baseline miss distance. A subject that follows the "sophisticated" rule best responds to a past self that guesses uniformly at random over the set of 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. [36]

unclassified participants ($p < 0.001$).

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

[36]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

4.2.2 Performance by number of modal type guesses

As in the *Replicate* treatment, we can consider how often a subject plays her modal type, i.e., her most frequently played behavioral type. This allows for a more continuous measure of Phase I behavior and lets us investigate the further robustness of our previous finding that classified subjects are better at best responding to their behavior than unclassified subjects. Figure 6 shows that making more modal type guesses in Phase I translates to more best responses in Phase II. A regression of the number of best-responses in Phase II on the number of modal type guesses 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 quite impressively shows the existence of subjects who are very good at best-responding to their Phase I guesses but have few modal type guesses. These subjects appear to represent omitted types. The fact that they so precisely best-respond to their guesses suggests that they indeed play omitted rules; they do not seem to 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 makes 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 find 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 classifications. 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 non-strategic 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.

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 ($p < 0.001$). 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 ($p < 0.001$).

There are, however, some omitted types (unclassified best responders), which suggests there is some room for additional strategic behavioral models. Note that the level k model captures about two-thirds of classified behavioral types, and behavioral types (including the equilibrium type), represent about two-thirds of "strategic" subjects. So, even if all strategic omitted types could be explained by a single new behavioral model, that new model (in our data) would capture fewer participants than does level k. In other words, our data predict that a new model cannot be as successful as level k in capturing strategic types in two-player guessing games.

4.3 Are Many Deterministic Rules Strategic?

The goal of this section is to assess the extent to which 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 that the subject also 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' exact Phase I guesses. Therefore, subjects who use rules of thumb may be able to replicate their guesses but fail to make the strategic leap that is 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 replications, they additionally must calculate the best responses. Note, however, that we find that subjects make virtually no mistakes when computing the best responses to guesses. As noted in Section 3, out of 600 times that subjects are tasked with responding to shown guesses, all but 6 are 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 (in Phase I). We saw that 72% of classified subjects are replicators and 81% are 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 ($p = 0.928$).[37]

[37] 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,

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.[38] 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.[39]

	Replicate	*BestRespond*	t-test
Classified Subjects (N)	18	26	
Miss Distance	49.13	53.54	0.750
Unclassified Subjects (N)	45	50	
Miss Distance	74.28	97.90	0.021

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 find that 39% are replicators, while only 20% are best-responders ($p = 0.046$). On average, unclassified subjects best-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.[40]

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%.[41]

$p = 0.964$.

[38]Furthermore, they realize about the same gains towards optimal play relative to the sophisticated baseline (see (Soph-Miss Dist.)/Soph). The sophisticated rule has a miss distance of 137.58 in the *Replicate* treatment and a miss distance of 105.03 in the *BestRespond* treatment, $p = 0.000$. For (Soph-Miss Dist.)/Soph) the numbers are 0.716 and 0.687, respectively, $p = 0.656$.

[39]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.

[40]In Phase I of the *BestRespond* treatment, unclassified subjects average 5.26 behavioral type guesses and 3.98 modal-type guesses, which is not significantly lower than the corresponding 4.89 ($p = 0.464$) and 3.71 ($p = 0.480$) measures from the *Replicate* treatment.

[41]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. The sophisticated rule has a miss distance of 135.81 in the *Replicate* treatment and of

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, suggests that 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 they reapply their original deterministic rules to recompute their former guesses. However, it could be that some participants merely have good numerical memories. In the *Memory* treatment, we provide a benchmark for how easy it is to remember 20 guesses that do not result from any deliberate rule. To this end, in Phase I of the *Memory* treatment, participants play the guessing games against computers that make random guesses. Notably, a subject sees her computer's guess before 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 (of course, without being shown these guesses).

As in our other treatments, we use a 0.5 unit window to determine whether or not an individual "remembers" a guess; we say a subject is a "rememberer" if she remembers 8 or more Phase I guesses. In Phase II of the *Memory* treatment, subjects remember between 1 and 7 guesses, so we find no "rememberers". On average, subjects remember 3.9 guesses out of 20. To assess that number, we compute the expected number of guesses that a subject with no numerical memory should be able to remember if she only recalls that, in Phase I, she best responded to her computer that chose a guess uniformly at random over the action set. For each game, this reasoning generates a unique action in Phase II that maximizes expected earnings. If all 20 subjects would have used this scheme, they would remember 2.8 guesses on average. Note that while not much smaller, this is significantly different than the mean of 3.9 remembered guesses ($p = 0.03$).[42]

Finally, we can compare how well subjects perform in the *Memory* and *Replicate* treatments. While 31 out of 63 subjects replicate 8 or more guesses and hence are replicators, no subject in the *Memory* treatment remembers 8 or more guesses ($p < 0.01$). The mean number of remembered guesses in the *Memory* treatment (3.9) is also significantly lower than the

114.01 in the *BestRespond* treatment ($p = 0.000$). 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. (Soph-Miss Dist.)/Soph has a value of 0.569 in the *Replicate* and of 0.449 in the *BestRespond* treatment ($p = 0.003$).

[42]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.

average number of replicated guesses (8.5) in the *Replicate* treatment ($p < 0.01$). Subjects in the *Memory* treatment even remember fewer guesses than unclassified subjects replicate (7.24) in the *Replicate* treatment ($p = 0.001$). A comparison to the *BestRespond* treatment yields similar results.[43]

In sum, participants who perform well Phase II of the *Replicate* and *BestRespond* treatments are unlikely to be exceptional in numerically remembering arbitrary Phase I play. Rather, their Phase II success likely comes from reimplementation of deterministic Phase I rules.

4.5 Explaining Unclassified Subjects

Our experimental design allows us to identify subjects who can replicate and best respond to their past behavior, even when we fail to capture that initial behavior with existing models. Such subjects are prime candidates for having deliberate rules that govern their Phase I choices. In this section, we use two methods aimed to shed light on the behavior of the 28 omitted types: the 18 unclassified replicators and 10 unclassified best-responders. Specifically, we aim to address whether many omitted types could be explained by one or a small number of models that may share the recursive best-response feature of level-k.

4.5.1 Method #1: Search for Types Related to $\{L1, L2, L3, EQ, D1, D2\}$

The $L1, D1$ and $D2$ strategies all involve best responding to a range of opponent guesses. Specifically, if a subject i performs 0, 1 or 2 rounds of iterative deletion of dominated strategies (*idds*) and believes her opponent plays uniformly at random among her remaining actions, then i's best response is the $L1, D1$ or $D2$ guess, respectively. The same $L1, D1$ and $D2$ actions emerge, however, as the best responses to the deterministic strategy of opponents guessing the *midpoints* of their guessing ranges that remain after, respectively, 0, 1 or 2 rounds of *idds*. Accordingly, we check whether there are subjects who perform 0, 1 or 2 rounds of *idds* and then guess their own midpoints at least 40% of the time.

Only one unclassified subject can be assigned to one of these three "midpoint types"; this participant (subject 80) performs one round of *idds* and guesses the midpoint of her remaining guesses. Furthermore, this subject is a best responder. Thus, when considering each deterministic strategy $s \notin BT = \{L1, L2, L3, EQ, D1, D2\}$ for which some $s' \in BT$ is a

[43] In the *BestRespond* treatment, 31 out of 76 subjects are 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 are even worse at remembering guesses than unclassified subjects are at best-responding to them in the *BestRespond* treatment, as they average 5.46 best-response guesses compared to 3.9 remembered guesses out of 20 ($p = 0.07$).

best response to s, we find one subject playing such an s with $D1$ as the corresponding s'. This subject is also one of our 28 omitted types.

Other strategies that are related to $BT = \{L1, L2, L3, EQ, D1, D2\}$ are $L4$, $BRD1$ and $BRD2$, the types who best respond to beliefs of $L3$, $D1$ and $D2$ opponents, respectively. (The best responses to $L1$, $L2$ and EQ are in BT already.) We do not find any unclassified subjects who can be explained by $L4$, $BRD1$, or $BRD2$.

4.5.2 Method #2: Search for Participants with Related Behavior

Thus far, we have checked three ways in which the 150 subjects from the *Replicate*, *BestRespond* and *ShowGuesses* treatment may be affiliated with a given type $t \in BT$. First, we checked whether a participant follows t, in which case the subject is classified. Second, we checked if an unclassified participant follows t', the determinisitc strategy to which t is a best response to; this occurs for one subject (Subject 80), who is an omitted type. Third, we checked if an unclassified participant follows a type t'' generated by taking the best response to type t, where t'' is not in BT; we find no such subjects. This threefold approach classifies 46 subjects, 45 in BT and one outside of BT who is one of our 28 omitted types.

To shed light on the behavior of the remaining 27 omitted types and address whether they employ strategies that are similar or related to each other, we follow an approach of CGC. We examine all pairs of subjects from the 150 participants in the *Replicate*, *BestRespond* and *ShowGuesses* treatment, and check for each pair of subjects i, i' whether their actions are similar. There are 4 asymmetric games that i and i' each play both as Player 1 and Player 2. For these games, we compare their guesses made as Player 1 as well as their guesses made as Player 2 (for a total of 8 guess-comparisons). For the 12 games (11 of which are asymmetric) that i and i' play only once, we compare i and i''s guesses in each game, even if i plays as Player 1 and i' plays as Player 2. If there are 8 or more guess-comparisons (out of 20) that are no more than 0.5 units apart, we say that i can be classified as an i'-type (and that i' can be classified as an i-type).

In addition to finding strategies shared by subjects and specifically by omitted types, we also want to check whether omitted types play according to a model that shares the recursive best response feature of level-k. Therefore, for each subject j, we define a br(j)-type by taking each j-type guess and computing its best response. If a subject j' has 8 or more guesses that are no more than 0.5 units apart from the br(j)-type, we say that j' can be classified as a br(j)-type.

We place two subjects h and h' into the same "cluster" if h' can be classified as an h-type or as a br(h)-type. One test whether such a clustering technique is helpful in finding subjects who play either according to the same strategy, or according to strategies that share a recursive best

response feature is to check whether it clusters subjects we have identified as either playing the same strategy or as playing strategies that best respond to each other, namely subjects classified as a type in BT.

We thus we hope to find at least three clusters of individuals: one linking the $D1$ subjects, another connecting the EQ players, and a large one that ties all $L1$, $L2$ and $L3$ participants together.[44] In fact, we find one large cluster, henceforth the *Behavioral Types Cluster (BTC)*, that contains, among others, 42 of our 46 classified participants. Let BTC_{42} denote these 42 classified participants who are linked in BTC. The four classified subjects that are not part of the BTC cluster are each of a different type: $D1$ (subject 24), EQ (subject 86), $L2$ (subject 61) and (the recently added) subject 80 to whom $D1$ is the best response.[45] Furthermore, each of these four subjects is in her own unit-sized cluster.[46]

In addition to linking 42 classified subjects, BTC captures 14 unclassified participants of which 6 are omitted types: Four are replicators (subjects 10, 28, 42 and 49) and two of are best responders (subjects 84 and 104). Subjects 42, 84, 104 are linked to BTC because each can be classified as an i-type for some $i \in BTC_{42}$. Subject 42 guesses her lower bound (which is often the equilibrium action) in 10 games and is a lower bound type. She is linked to Subject 23, an EQ player in BTC_{42}. Subjects 84 and 104 are linked to a number of L1 subjects in BTC_{42} because, in 11 and 17 games, respectively, their guess is within 0.5 units of *their opponent's* unbounded $L1$ guess, that is they guess $R_i(t_j[l_i + u_i]/2)$.

Subjects 10, 28 and 49 are linked to BTC because each is best responded to by some $i \in BTC_{42}$. Subjects 49 and 10 (both replicators) are best responded to by EQ subjects in BTC_{42} because even though they are not an EQ type (though subject 10 has five EQ guesses) their guesses often fall within the range of guesses that yield EQ as their opponents' best response.[47] Subject 28 is best responded to by three $L2$ individuals in BTC_{42}. Upon investigating her guesses one-by-one, we find she is actually a very deliberate player: in 19 games, she guesses as close as possible to the product of her midpoint with her target, $R_i(t_i[l_i + u_i]/2)$. To see how the $L2$ individuals best respond to subject 28, note that the "unbounded" version of this strategy is $t_i[l_i + u_i]/2$, which, when multiplied by t_j, yields $t_j t_i[l_i + u_i]/2$, which is the

[44] These clusters are not guaranteed to arise; of our 45 classified participants, only three subjects, subject 36, 102 and 137 implement their behavioral types in all 20 games.

[45] Of those four behavioral types that are not in the BTC cluster only two: subject 86 and the newly discovered subject 80 are best repsonders. The other two are in the *Replicate* treatment and fail to replicate.

[46] Recall that the cluster analysis does not necessarily link two subjects classified as the same $t \in BT$. For instance, subject i and j may guess the Level 1 actions in the 11 asymmetric games that they play as Player 1 and as Player 2, respectively, and otherwise, play randomly. Each will be classified as L1 yet may have no guesses that are within 0.5 units of one another, and hence, will not be part of the same cluster.

[47] These two subjects are also not lower (upper) bound types, i.e. they fail to guess within 0.5 of their lower (upper) bounds at least 40% of the time. Subject 49 is best responded to by an EQ player, while subject 10 is best responded to by an EQ player as well as by an $L2$ subject.

unbounded version of $L2$.

Therefore, the six omitted types related to subjects in BTC_{42} have decision rules that are similar to the strategies in BT. Nevertheless, their behavior is generally rather unconventional (such as guessing as close as possible to the opponent's $L1$ guess).

The largest cluster next to BTC contains three subjects: subject 25 and 56 who are replicators, and subject 93 who is not a best responder. Subjects 56 and 93 are linked to each other, as both make 10 guesses within 0.5 units of the rather bizarre strategy given by $R_i([R_i(t_i l_j) + R_j(t_j u_i)]/2) \equiv s_{93}$. In words, an s_{93} type guesses as close as possible to the average of her lowest undominated guess and her opponent's largest undominated guess. Subject 25 makes 9 guesses within 0.5 units of the actions given by $R_i([t_i u_j + t_j l_i]/2) \equiv s_{25}$. In words, an s_{25} player guesses as close as possible to the midpoint of x and y, where x is the product of her own target and her opponent's upper bound and y is the product of her opponent's target and her own lower bound.[48]

There are two additional clusters bigger than size one that contain an omitted type. One contains subjects 53 (a replicator) and subject 136 who is not a best responder. Each computes her own L1 guess as well as that of her opponent and guesses the average, i.e., follows the rule given by $R_i(1/2 \times [R_i(t_i(l_j + u_j)/2) + R_j(t_j(l_i + u_i)/2])$. Subjects 53 and 136 have 11 and 16 such guesses within 0.5 units of this rule, respectively.

The other cluster consists of subject 112 who is a best responder and subject 139 who is not. Each makes 10 lower bound guesses. Regarding the 11 asymmetric games played only from one side, these subjects play each from the same side. Subject 42, a previously discussed lower boundary type who is part of BTC, plays these games from the other side.

The remaining individuals are all in their own unit-sized clusters, save for a cluster of two, a non-replicator (subject 13) and a subject from the *ShowGuesses* treatment (subject 67). That is 17 of the 27 omitted types are in these unit-sized clusters which makes it difficult to find their strategies.

Thus, to summarize, the cluster analysis reveals a sizable BTC cluster that includes 42 of the initially classified subjects as well as 14 unclassified participants of which 6 are omitted types. Outside BTC, there are 18 omitted types who are linked to no other subjects, though one omitted type (subject 80) is classified as using a strategy that is connected to BT, see Section 4.5.1. The four omitted types who are part of clusters with other subjects are spread thinly across three different clusters of sizes two and three. Taken altogether, these results

[48] At first, s_{25} appears quite different from s_{93}. Subject 25 is linked directly to subject 93 and the reason is as follows. Suppose that a player follows a "less bounded" version of s_{93}, namely, $R_i([t_i l_j + t_j u_i]/2) \equiv s_{93'}$. In many games, s_{93} and $s_{93'}$ yield the same action. Upon inspection, the portions inside the square brackets of $s_{93'}$ and s_{25} are identical except for the i and j roles being flipped. Thus, subjects 25 and 93 are linked through making the same guesses in asymmetric games.

cast doubt over the possibility that an alternative model of a unique deterministic strategy, or even an alternative model that shares the recursive best response feature of level-k could come close to describing as many individuals as the level-k model (which classifies 26 subjects in our experiment).

5 Related Literature, Methodological Remarks, Discussion and Gender

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, does not guarantee that play is predictable on an individual level. 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. (2015) for a correlation of play with a CRT measure and Agranov, Caplin and Tergiman (2015) 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. Most research on stability of rules within individuals has focused on comparing behavior across strategic settings.[49] Crawford and

[49] 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,

Iriberri (2007) look at various hide-and-seek games and find some consistency across games. On the other hand, Buchardi and Penczynski (2011) and Georganas, Healy, and Weber (2015) 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.[50] Indeed, our results from the *BestRespond* treatment suggest that level-k subjects are in general not rule of thumb players. There are several other 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 (2015) (see also Georganas et al. (2015) below).[51]

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. (2014), 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 (2014), 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

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.

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

[51] 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 (2015) 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.

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 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.[52] Alternatively, researchers have manipulated beliefs to determine whether the behavior of subjects changes accordingly. Georganas et al. (2015) 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

[52] 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).

Replicate and *BestRespond* treatments then help elucidate the subjects' processes of strategic choice.

The main advantage of the approach we take in this paper is that if subjects are playing according to a behavioral game theory type (or indeed any deterministic rule), 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.

5.1 Gender

The rising interest in gender differences in economics (see e.g. Niederle, 2016) has also lead to an interest on whether women are less strategic than men. For an early example see Casari et al. (2007). They find that women, compared to men, are much more prone to bid above the risk neutral Nash equilibrium in common value auctions. However, as women and men gained experience, their bidding behavior converged.

In this paper we can use such a test common to the literature on whether men are more "sophisticated" than women by assesing whether women or men have higher earnings in Phase I of the *Replicate*, *BestRespond* and *ShowGuesses* treatment. Of our 150 subjects 72 are female. The average expected earnings per game are \$1.01 for women and \$1.04 for men ($p = 0.172$). When we regress expected earnings per game on a gender dummy and also control whether the subject sees the games from the point of view of Player 1 (P1) or Player 2 (P2), we find that women have no more than 3 percent lower earnings than men, a difference that is not significant, see Table 5.[53]

One problem with such an endeavor is that different actions can often be rationalized by different beliefs. A more direct and more stringent test is therefore to focus on actions that cannot be rationalized, namely dominated guesses. The average number of dominated guesses are 2.65 for women and 2.06 for men ($p = 0.180$). A regression controlling for the point of view from which the subject played confirms this result, see Table 5.

[53]Instead of focusing on earnings, we can focus on whether subjects behave according to a behavioral strategy: On average in Phase I, men and women have 6.49 and 6.26 modal type guesses ($p = 0.769$). Twenty five out of 78 men and 20 out of 72 women are classified in Phase I. These proportions are not significantly different using a two-tailed Fisher's exact test ($p = 0.597$).

	Phase I		Phase II Earnings		
	Earnings	Dominated Guesses	*Replicate*	*BestRespond*	*Replicate & BestRespond*
	(1)	(2)	(3)	(4)	(5)
Female	-0.03	0.55	-0.12*	-0.01	-0.06
	(0.02)	(0.44)	(0.06)	(0.06)	(0.04)
Player 1	0.04*	-0.52	-0.01	0.09	0.04
	(0.02)	(0.44)	(0.06)	(0.06)	(0.04)
BestRespond					-0.24***
					(0.04)
Constant	1.02***	2.34***	1.65***	1.31***	1.60***
	(0.02)	(0.38)	(0.05)	(0.06)	(0.05)
Observations	150	150	63	76	139

TABLE 5.—Phase I and II Earnings are in dollars. Phase I Dominated Guesses are the total for Phase I. The table shows the results of linear regressions with standard errors are clustered at the individual level and shown in parentheses. *Female* is a gender dummy equal to 1 if the subject is female. *Player 1* is a dummy equal to 1 if the subject is in group P1 (see Footnote 5). *BestRespond* is a dummy equal to 1 if the subject is from the *BestRespond* treatment. Regression (3) includes the 63 subjects (out of 64) from the *Replicate* treatment whose computers functioned properly in Phase II. Regression (5) includes the subjects from Regressions (4) and (5). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

While women do not make significantly more dominated guesss than men, our design allows us to asses whether women are more or less likley to deliberately use a well-defined strategy rather than making idiosyncratic choices that are unlikley to be driven by a deterministic rule. We have two direct tests to assess such gender differences in "sophistication". We can assess whether women are more or less able to replicate their guesses and hence whether they are more or less likley to use a well-defined deterministic rule than men are. Second, we can assess whether women are more or less strategic, that is, whether they are more or less able to best respond to their guesses than men are.

Women seem to be slightly less good at replicating their guesses than men, their earnings per game in Phase II of the *Replicate* treatment are about 12 cents lower per game than those of men, a significant difference (see Table 5). However, this difference may be driven by the 31 women in the *Replicate* treatment, as they also earn about 5 cents less per game than men in Phase I, a difference not bourne out when we consider all 72 women in the *Replicate*, *BestRespond* and *ShowGuesses* treatment.[54] Note, however, that the expected earnings in Phase II of the *Replicate* treatment are not affected by Phase I earnings, as they only depend on how close the Phase II guess is to the Phase I guess, a distance that is minimized (and zero) by replicating the Phase I guess, whatever that may be.

When we assess whether women are less strategic than men, we compare the Phase II earnings in the *BestRespond* between women and men. Women earn only about 1 cent less than men per game, a difference that is not significant, see Table 5.[55] Finally, we can combine the Phase II earnings of both treatments, and find once more that women earn about 6 cents less than men per game, a difference that is not significant (see Table 5). As a comparison, subjects earn \$1.59 and \$1.35 in Phase II of the *Replicate* and *BestRespond* treatments, a significant difference ($p < 0.01$). Using our very direct test of strategic sophistication we cannot confirm that women are less strategically sophisticated than men.

6 Conclusion

To date there has not been a practical way to organize players according to whether they implement deliberate decision rules, especially if we haven't behavioral models to explain their behavior. In this paper we say that a subject deliberately employs a well-defined rule if the

[54] A linear regression of expected Phase I earnings per game of the 31 women and 33 men in the *Replicate* treatment yields a constant of 1.04 (s.e. $0.026, p < 0.01$), a coefficient on a female dummy of -0.05 (s.e. $0.030, p = 0.095$) and a coefficient on a Group 1 dummy of 0.04 (s.e. $0.030, p = 0.197$).

[55] In Phase II of the *BestRespond* treatment we can also once more assess whether women are more likely to make dominated guesses than men do. This is not the case: a linear regression on the number of dominated guesses of the 37 women and 39 men yields a constant of 1.66 (s.e. $0.393, p < 0.01$), a coefficient on a female dummy of 0.24 (s.e. $0.481, p = 0.616$) and a coefficient on a Group 1 dummy of -0.111 (s.e. $0.481, p = 0.818$).

behavior of the subject conforms to an expected relationship across strategic situations. We provide an environment and a test that allow us to identify such behavior, enabling 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 each subject's optimal action depends solely on her own previous choices. Subjects' behavior 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.

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, 2014, Enke and Zimmermann, 2015 and Vespa and Wilson, 2016). 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.

Finally, we show how having participants replicate their past choices can be a method to assess whether a subject deliberately uses deterministic strategies versus makes choices idiosyncratically. This approach may not detect the deliberate use of a mixed strategy, however. We think our method can be extended to mixed strategies, albeit probably in simpler contexts. For instance, consider the case where a mixed strategy may be optimal, as in a matching pennies or two-player zero sum game. One could provide participants with a randomization device and have subjects decide how to randomize. One could then show subjects the same game again and ask them to randomize in the same way they did previously. We leave it to future research to explore our method in settings that are more likely to invoke the use of mixed strategies.

7 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. 2015. "Naive Play and the Process of Choice in Guessing Games" *Journal of Economic Science Association*, Vol. 1(2), pp. 146-157.

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. 2014. "Imperfect Choice or Imperfect Attention? Understanding Strategic Thinking in Private Information Games." *Review of Economic Studies*, 81(3), 944-970.

Burchardi, Konrad B., and Stefan P. Penczynski. 2014. "Out Of Your Mind: Eliciting Individual Reasoning in One Shot Games." *Games and Economic Behavior* 84: 39-57.

Burnham, Terence C., David Cesarini, Magnus Johannesson, Paul Lichtenstein, Bjorn Wallace. 2009. "Higher cognitive ability is associated with lower entries in a p-beauty contest," *Journal of Economic Behavior and Organization*, 72, 171-175.

Camerer, Colin F., Teck-Hua Ho, and Juin-Kuan Chong. 2004. "A Cognitive Hierarchy Model of Games." *Quarterly Journal of Economics* 119 (3): 861-98.

Casari, Marco, John C. Ham and John H. Kagel. 2007. "Selection Bias, Demographic Effects, and Ability Effects in Common Value Auction Experiments." *American Economic Review*, 97(4): 1278–1304

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): 9163-68.

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, 5-62.

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, 299-314.

Enke, Benjamin and Florian Zimmermann, "Correlation Neglect in Belief Formation," working paper, 2015.

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. 2014. "Hypothetical Thinking and Information

Extraction in the Laboratory," *AEJ: Microeconomics*, 6.4, 180-202

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. 2015. "On the Persistence of Strategic Sophistication." *Journal of Economic Theory*, 159(A): 369400.

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 Winner's 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, 1313-1326.

Niederle, Muriel. 2016. "Gender", The Handbook of Experimental Economics, vol 2. Editors, John H. Kagel and Alvin E. Roth. Princeton University Press.

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): 572-85.

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.

Vespa, Emanuel and Alistair Wilson. 2016. "Communication with Multiple Senders: An Experiment," *Quantitative Economics* 7.1, 1-36.

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.

8 Appendix

8.1 Classified Subjects: Comparison with CGC

We found that 30% of participants are classified using the apparent type method. While we use the same apparent type classification method as CGC, they have significantly more subjects classified as behavioral types, 49% ($p = 0.005$). Table 6 shows the fraction of participants classified as each behavioral type. Most notably we have fewer $L1$ and $L2$ types, though roughly the same number of equilibrium types.

There are two potential reasons why we have a different number of subjects classified as behavioral game theory types using the apparent type method than CGC has. The first concerns the games we use and the second the subjects.

We say that a game has type separation of K for player i, if for any types $\tau_i^1, \tau_i^2 \in \{L1, L2, L3, L4, EQ\}$ with $\tau_i^1 \neq \tau_i^2$ we have $|\tau_i^1(x) - \tau_i^2(x)| \geq K$, where $\tau_i^j(x)$ is the action prescribed by strategy τ_i^j for $j = 1, 2$. In our experiments, subjects play 8 random games, that have a type separation of at least 30, games $11 - 18$ in Table 1, and 4 of the CGC games that have type separation of at least 10, games $1 - 6$ in Table 1, of which they play two from both sides. This results in 14 games with type separation, or 70% of all games. In contrast, of the 8 CGC games only 4 have type separation. Since in CGC subjects play every game from both sides, this results in 50% of games with type separation.

To assess the role of the type of games for the classifications of participants, we make two comparisons. First, we compare the classification in all 20 games to the classification we would have obtained had we only used the 14 games with type separation (see Type-Sep Games in Table 6). For all comparisons we keep a threshold of 40% for the apparent type classification. That is, a subject has to have a guess not further than 0.5 from the same behavioral game theory type guess in at least 40% of games to be classified as that behavioral type. The number of classified subjects drops from 30% to 24% when we go from using all 20 games to only using the 14 games with type separation.

Second, since our subjects play 10 CGC games and 10 new games, two of which have a dominant strategy for one player, we can compare the classification in those two subsets of games, that have 60% and 80% of games with type separation, respectively. While 39% of subjects are classified in the 10 games we use from CGC, only 33% of subjects are classified in the new games. While almost a 16% drop, this difference is not significant, $p = 0.335$. Note, however, that the number of subjects classified in just the CGC games in our data is not significantly different from the classification in the CGC data (92 unclassified subjects out of 150 is not significantly difference from 45 unclassified subjects out of 88, $p = 0.137$). When we compare the number of subjects unclassified in the 10 non-CGC games (101 out of 150) to

the CGC data (45 out of 88), the difference is significant, $p = 0.019$.

A second possible explanation is that our participants are not as sophisticated as the students used by CGC, or that they are not sufficiently motivated given the incentives at hand. Recall, however, that we have 20 participants in the *Memory* treatment who in Phase I best respond to the guess of a computer they observe, and 10 participants in the *ShowGuess* treatment who in Phase II best responded to their Phase I guess *after* observing their Phase I guess. Of the 400 guesses made by the 20 participants in the *Memory* treatment, all but 3 are within 0.5 of the best response, and of the 200 guesses made in the *ShowGuess* treatment, all but 2 are within 0.5 of the best response. This suggests that our participants are willing and able to calculate the best response to a guess, even when we only pay them, as in these two cases at most \$1 per guess.

To assess the sophistication of subjects we can also assess the extent to which they make dominated guesses which are not accounted for by any behavioral game theory model. Only about one third of subjects (44) have no dominated guess, though two thirds (97) have two dominated guesses or less. While CGC do not have a similar analysis they have more exclusion criteria than we do. This way they may eliminate players who make many mistakes.

When we condition only on participants that have no dominated guess, we have 52% of participants who are classified. This is a significantly higher fraction than the 21% of those subjects who have at least one dominated guess, $p < 0.01$.[56] We can compare the fraction of subjects who are classified among participants with no dominated guess between our data and CGC. The difference is still significant ($p = 0.069$).

8.2 Apparent type classifications using different parameters

In the following two figures we show the relative distribution of types as a function of various cutoffs. For Figure 7, we keep a 0.5 ball around the behavioral type guess. 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. Since the modal type may not be unique, we count a subject that has n behavioral strategies $\{m_1, ..., m_n\}$ for her modal type as $1/n$ of an m_i player, for $m_i \in \{L1, L2, L3, EQ, D1, D2\}$ for $1 \leq i \leq n$. For any $q \in \{1, ..., 20\}$ and any behavioral type m_i, we can compute the number of subjects whose modal type corresponds to m_i and who match that type in q or more games. Figure 7 shows for each number of games q, for each behavioral type $m_i \in \{L1, L2, L3, EQ, D1, D2\}$, the number of subjects who in at least q games

[56] When we condition on participants with one or two dominated guesses, we have 25% of participants who are classified, significantly lower than the 52% who had no dominated guess ($p = 0.006$). The number is, however, roughly similar to the 17% who are classified among participants with three or more dominated guesses ($p = 0.473$).

TABLE 6.—Classification comparison with CGC

ALL SUBJECTS	$L1$	$L2$	$L3$	EQ	$D1$	$D2$	Uncl.	N
Our Data								
All Games (20)	9.3%	6.7%	1.3%	10%	2.7%	0%	70% (105)	150
Type-Sep Games (14)	6.7%	4.7%	1.3%	9.3%	2%	0%	76% (114)	150
Just CGC Games (10)	13%	8%	1.7%	12%	3%	1%	61.3% (92)	150
Non CGC Games (10)	9.3%	8.3%	1.3%	11.3%	2.3%	0%	67.3% (101)	150
CGC Data								
All Games (16)	22.7%	13.6%	2.3%	10.2%	0%	0%	51.1% (45)	88
Our Data: Rational								
All Games (20)	15.9%	9.1%	2.3%	20.5%	4.5%	0%	47.7% (21)	44
Type-Sep Games (14)	13.6%	4.5%	2.3%	13.6%	4.5%	0%	61.4% (27)	44
Just CGC Games (10)	19.3%	11.4%	3.4%	22.7%	5.7%	3.4%	34.1% (15)	44
Non CGC Games (10)	15.9%	11.4%	2.3%	15.9%	4.5%	0%	50% (22)	44
CGC Data: Rational								
All Games (16)	27%	21.6%	5.4%	18.9%	0%	0%	27% (10)	37

For several subsets of games the fraction of participants classified as various types (or left unclassified). We compare data from this paper (Our Data) to data from CGC (CGC Data). "Type-Sep" Games refers to the 14 of our 20 games that have type separation of at least 10.5, "Just CGC Games" refers to the 8 games from CGC used in our experiment, of which 2 were played from both sides, "Non CGC Games" refers to the 8 randomly drawn games with type separation of at least 30 and the 2 games with a dominant strategy. The part of the Table with the heading "Rational" refers to analyses where we only include subjects that made no dominated guess in any game of the experiment.

play m_i – up to 0.5 – and who have m_i as their modal type. When we require subjects to play the same behavioral type in only one game in order to be classified all but 2 of the 150 subjects are classified. While $L1$ and EQ are the most common types, $L1$ is more prevalent when we require subjects to play a type only in 6 games or less to be classified. Once the threshold is 7 or more games (up to 13 or less), EQ is slightly more prevalent. However, overall the figure shows that the relative distribution of types is quite stable.

Figure 8 shows that similar conclusions hold when we allow subjects to deviate up to 5 instead of 0.5 from each behavioral type guess.

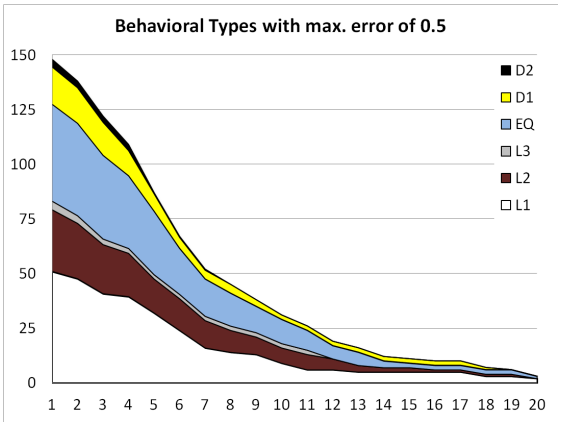

FIGURE 7.—The number of subjects classified as a specific behavioral type when we require subjects to play at least q games with a guess at most 0.5 different from that behavioral type guess to be classified.

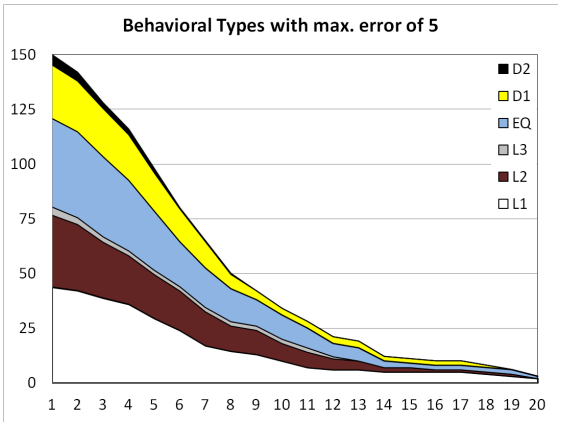

FIGURE 8.—The number of subjects classified as a specific behavioral type when we require subjects to play at least q games with a guess at most 0.5 different from that behavioral type guess to be classified.

8.3 The number of behavioral type guesses dependent on the number of modal type guesses

For each participant we compute the modal type (the behavioral type they use most often), and compute the number of modal type guesses made. We then compute the total number of behavioral type guesses made. This will be, of course, at least as large as the number of modal type guesses. It may be larger if a subject switches between several behavioral types. Figure 9 shows that 91% of subjects have at most only 3 behavioral type guesses that are not their modal type.

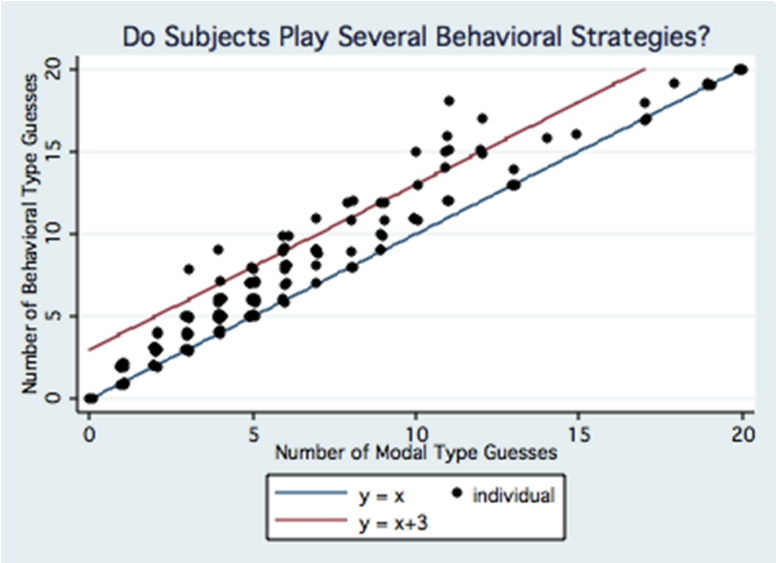

FIGURE 9.—For each number of modal type guesses of a subject, the number of total behavioral type guesses that subject made.