

# ON THE PERSISTENCE OF STRATEGIC SOPHISTICATION<sup>†</sup>

SOTIRIS GEORGANAS\*, PAUL J. HEALY\*\* AND ROBERTO A. WEBER\*\*\*

**ABSTRACT.** Levels-of-reasoning models have been used to interpret behavior in laboratory games. We test whether these models generate reliable cross-game testable predictions. Within one family of similar games subjects' observed levels are fairly consistent, but within another family of games there is virtually no cross-game correlation. Moreover, the relative ranking of subjects' levels is not consistent across games. Direct measures of strategic intelligence are not correlated with observed levels of reasoning. We conclude that if strategic sophistication is a persistent trait of a person, it is not identified by their observed level in this model.

**Keywords:** Level- $k$ ; cognitive hierarchy; behavioral game theory.

**JEL Classification:** C72; C91; D03.

Draft: March 18, 2010

## I INTRODUCTION

Following a considerable body of literature demonstrating deviations from Nash equilibrium play in experimental games (see, for example, Camerer, 2003), behavioral research has sought to model the processes determining individual play and aggregate behavior. One widely-used approach for modeling behavioral deviations from Nash equilibrium in one-shot games involves the use of heterogeneous types, based on varying levels of strategic sophistication (Nagel, 1993; Stahl and Wilson, 1994; Costa-Gomes et al., 2001; Camerer et al., 2004).<sup>1</sup> In this framework, often referred to as “Level- $k$ ” or “Cognitive

---

<sup>†</sup>The authors thank Tilman Börgers, Colin Camerer, John Kagel, Stephen Leider, John Lightle, Tom Palfrey, Reinhard Selten, Dale Stahl, and Joseph Tao-yi Wang for their valuable comments.

\*Dept. of Economics, Royal Holloway, University of London, Egham, Surrey TW20 0EX, England; sotiris.georganas@rhul.ac.uk

\*\*Dept. of Economics, The Ohio State University, 1945 North High street, Columbus, OH 43210, U.S.A.; healy.52@osu.edu.

\*\*\*Dept. of Social & Decision Sciences, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, U.S.A.; rweber@andrew.cmu.edu.

<sup>1</sup>An alternative approach involves modeling deviations from Nash equilibrium as noise (or unobservable utility shocks) in players' best response. For an example, see the Quantal Response Equilibrium model proposed by McKelvey and Palfrey (1995). Rogers et al. (2009) bridges the Quantal Response approach with the “Level- $k$ ” approach studied here. Other directions in behavioral game theory include the study

Hierarchies”, players’ strategic sophistication is represented by the number of iterations of best response they perform in selecting an action.

In the simplest version of such models, Level-0 types randomize uniformly over all actions and, for all  $k > 0$ , the Level- $k$  type plays a best response to the actions of Level- $(k - 1)$ . Thus, the model suggests that a subject’s type is a measure of her strategic sophistication, or more precisely, her belief about the strategic sophistication of her opponents. The application of such models to data from one-shot play in experiments has yielded several instances in which the model accurately describes the aggregate distributions of action choices. We provide a review of this literature in the next section.

While the value of the Level- $k$  framework as a *post hoc* descriptive model of the aggregate distribution of actions in laboratory game play has been widely documented, an open question remains regarding whether Level- $k$  types correspond to some meaningful individual characteristic that one might label as “strategic sophistication”. That is, does an individual’s estimated “level” correspond to a persistent individual quality that can be used to predict play across games? If levels are indicative of strategic sophistication then there should indeed exist reliable cross-game correlations in players’ levels. Moreover, if levels measure a reliable individual characteristic, the possibility exists to use estimated types to predict players’ behavior in novel games and economic contexts and to estimate types using measures other than actual game play.

In this paper, we explore whether some persistent individual characteristic emerges when one observes an individual player’s behavior across games. We begin by identifying several reasonable restrictions on cross-game behavior for a model of strategic sophistication that takes as its starting point the Level- $k$  framework. We then conduct a laboratory experiment in which subjects play several games, drawn from two distinct families of games, including a novel family of games that allow a straightforward classification of Level- $k$  types. Within each game, we identify an individual’s type in a Level- $k$  framework. We then evaluate the persistence of a subject’s type across games, as a way of identifying whether the Level- $k$  classification measures something akin to strategic sophistication. We test various ways in which Level- $k$  types may be persistent. For example, the most stringent definition of persistence requires a player’s level to be invariant across all games. A weaker notion requires only that players’ relative ranking of levels be invariant, so that an observation of one player playing a higher level than

---

of dynamics following initial play (see Crawford, 1995; Erev and Roth, 1998; Camerer and Ho, 1998, for example) or other-regarding preferences (Fehr and Schmidt, 1999, e.g.).

another in one game implies that the player will continue to play a higher level in all other games.

We also consider two additional ways in which strategic sophistication might be detectable. First, to further explore whether the Level- $k$  classifications are indeed correlated with some notion of strategic sophistication, we elicit several direct measures of strategic intelligence using brief quizzes that are known to identify strategic reasoning ability or general intelligence. We explore the relationship between such measures and subjects' Level- $k$  types identified from their behavior. Second, we have subjects play each game against three different opponents: a subject randomly selected from the population in the session, the subject who scored highest on the strategic intelligence measures discussed above, and the subject who scored lowest. Thus, we are able to detect whether sophisticated types vary their behavior based on the expected sophistication of their opponent.

The degree of persistence in strategic sophistication that emerges from our data is mixed and generally weak. Our key results are summarized as follows:

- (1) The *aggregate* distribution of types is similar to that found in previous studies, though the distribution varies substantially between games.
- (2) Levels are moderately persistent within one family of games, but not persistent within the other.
- (3) For any two players, the relative ordering of their levels is not particularly stable between games, especially across the two families of games.
- (4) The quizzes generally fail to predict players' levels, though Level-1 play is weakly correlated with a test for autism and poor short-term memory.
- (5) Some players adjust strategies against stronger opponents, but neither quiz scores nor levels predict which subjects make this adjustment.
- (6) When a game's payoff function is modified to introduce dominant strategies, which should be played by most types, players' behavior does not significantly change and most players do not select the dominant strategy.

Overall, we conclude that the Level- $k$  framework is helpful for understanding heterogeneous behavior within a class of games, as demonstrated by Stahl and Wilson (1994) or Costa-Gomes et al. (2001), but it does not easily map into a classification of subjects' underlying strategic sophistication.

## II LITERATURE REVIEW

The notion of heterogeneous strategic sophistication operating through limited iterations of best response dates back at least to the ‘beauty contest’ discussion of Keynes (1936). Nagel (1993) and Ho et al. (1998) (HCW) explore behavior in laboratory  $p$ -beauty contest games and describe the resulting distribution of levels according to the Level- $k$  model. In these games  $n$  subjects simultaneously select number choices from an interval and the winner is the subject whose guess is closest to  $p$  times the group average. Most initial guesses are far from equilibrium but generally conform to the first four levels of reasoning (Level-0 through Level-3) in the Level- $k$  model (see also Duffy and Nagel, 1997 or Bosch-Domènech et al., 2002.)

Stahl and Wilson (1994) study Level- $k$  behavior in ten  $3 \times 3$  games.<sup>2</sup> They find that roughly 25 percent of players are Level-1, 50 percent are Level-2, and 25 percent are Level- $\infty$  (Nash) players. Level-0 play is virtually non-existent. Stahl and Wilson (1995) examine play in twelve normal-form games played without feedback. They add to the Level- $k$  model a “Worldly” type who knows the other 4 types exist, but thinks he is unique in this knowledge, and a “Rational Expectations” type who knows about all other types and himself. They find little evidence of Level-0 and Rational Expectations types; the frequencies of the other types fall monotonically in the level.

Camerer et al. (2004) present a variation of the Level- $k$  model in which players best respond to the empirical distribution of levels truncated below their own level. Thus, a Level- $k$  player believes all other players are Level-0 through Level- $k - 1$  and his belief about the relative frequencies of those levels is accurate. Using a Poisson distribution of Levels reduces the model to a single parameter (after defining the Level-0 distribution) that describes the mean level in the population. They find their model fits experimental data well with an average level of around 1.6, and that more educated populations tend to exhibit higher average levels.

Costa-Gomes et al. (2001) fit an augmented Level- $k$  model to the behavior of subjects who play 18 games of varying difficulty. They allow for nine different types, including Level-1, Level-2, Altruistic (maximizing the sum of payoffs), Pessimistic (playing maxmin strategies), Optimistic (playing max-max strategies), Equilibrium, D1 (deleting opponents’ dominated strategies), D2 (using two rounds of deletion of dominated

---

<sup>2</sup>In their model Level-0 players are assumed to randomly choose strategies, Level-1 players best respond to Level-0, and Level-2 players best respond to a Level-1 strategy with noise added. This works similarly to best responding to a mixture of Level-0 and Level-1.

strategies), and Sophisticated (best responding to the empirical distribution of strategies). In their experiment, payoffs in the games are initially hidden to subjects, so that estimation of a player’s level based on strategy choice can be augmented by analyzing which pieces of information subjects choose to view before making a decision. They find mostly Level-1, Level-2, and D1 types and generally see more “strategic” types in simpler games. Their estimation based on choice behavior assumes each player mixes uniformly with probability  $\varepsilon$  and plays according to her type otherwise; the estimates of  $\varepsilon$  are fairly high, typically near 30%, indicating an imperfect fit with the model.

In Costa-Gomes and Crawford (2006) (CGC06) players participate in 16 two-person guessing games in which a player and her opponent are each assigned an interval  $[a_i, b_i]$  and a ‘target’  $p_i \in \{0.5, 0.7, 1.3, 1.5\}$ . Players’ payoffs decrease in the distance between their own guess and  $p_i$  times their opponent’s guess.<sup>3</sup> Again, lookup behavior is used to strengthen type estimation. The estimation is similar to the previous paper, adding uniform play with probability  $\varepsilon$  and logistic errors to each type’s strategy. Again the results support the Level- $k$  model: A reasonably large percentage of players play exactly the strategy predicted by one of the Level- $k$  types. Six of the ten games we study in this paper are two-person guessing Games; we compare our findings to CGC06 in the analysis below.

Crawford and Iriberry (2007a) analyze ‘hide-and-see’ games, which are expanded matching-pennies games with labeled strategies to induce focal effects. They find that a Level- $k$  model with Levels 0–4 fit the data well, though the assumption that Level-0 types favor focal strategies appears important for the model’s fit. Chen et al. (2009) study similar hide-and-see games on a two-dimensional grid. They use eye-tracking technology to augment the type estimation based on behavior alone. They find distributions of types that somewhat more uniform than in past studies. When subjects’ data is randomly re-sampled to generate new bootstrapped samples, only 8 of 17 subjects receive the same classification in at least 95% of the bootstrapped samples as they did in the original sample, suggesting that many subjects’ behavior is not strongly consistent with any one level across these games.

Finally, several recent papers apply the Level- $k$  concept to study departures from Nash equilibrium play in auctions. For example, Crawford and Iriberry (2007b) apply the Level- $k$  model to auction data from many different experiments and find that in many

---

<sup>3</sup>Two-person guessing games differ from (two-player)  $p$ -beauty contest games in that the former allows the players to have different intervals and targets, while the latter does not.

cases—though not all—Level- $k$  yields a significantly better fit than the Nash equilibrium. Georganas (2009) applies the Level- $k$  model to auctions with resale. Using logistic errors he finds it yields a much higher likelihood than the Nash model, as does a quantal response equilibrium. This result depends crucially on choosing a random level zero instead of a truthful one, under which the Level- $k$  model is observationally equivalent to a Nash equilibrium. However, Ivanov et al. (2008) use a clever design in which players in second-price common-value auctions bid against their own earlier-period strategies to demonstrate that overbidding (‘the winner’s curse’) cannot be explained by subjects’ misguided beliefs about their opponents as in the Level- $k$  framework.

### III A FORMULATION OF LEVEL- $k$ MODELS

The usual applications of the Level- $k$  model treat it as an *ex post* descriptive model and as such it lacks cross-game or cross-individual testable restrictions. In this section we lay down a formal framework in which such testable restrictions can be defined clearly. Our experiments then examine several possible cross-game testable restrictions to see which have empirical merit.

Specifically, we build a type-space model for two-player games where an agent’s type describes her *capacity* for iterated best-response reasoning and her realized *level* of iterated best-response reasoning. Under Harsanyi’s (1967) interpretation, types would also describe beliefs about opponents’ types, second-order beliefs about opponents’ beliefs, and all higher-order beliefs. Following the Level- $k$  literature, however, we make the simplifying assumption that a player’s level is a sufficient statistic for their hierarchy of beliefs, and that all players believe all others to have strictly lower levels than themselves.<sup>4</sup>

In our experiment subjects play several two-person games. Let  $\gamma = (\{i, j\}, S, u)$  represent a typical two-person game with players  $i$  and  $j$ , strategy sets  $S = S_i \times S_j$ , and payoffs  $u_i : S \rightarrow \mathbb{R}$  and  $u_j : S \rightarrow \mathbb{R}$ . The set of all such two-player games is  $\Gamma$ . When players use mixed strategies  $\sigma_i \in \Delta(S_i)$  we abuse notation slightly and let  $u_i(\sigma_i, \sigma_j)$  and

---

<sup>4</sup>For example, Costa-Gomes et al. (2001), Costa-Gomes and Crawford (2006), Crawford and Iriberry (2007b), and Crawford and Iriberry (2007a) assume that all players with a level of  $k > 0$  believe all other players’ level to be  $k - 1$  with probability one. Camerer et al. (2004), on the other hand, assume that all players with a level of  $k > 0$  believe the realized levels of his opponents to follow a truncated Poisson distribution over  $\{0, 1, \dots, k - 1\}$ . Whatever the assumption on first-order beliefs, all higher-order beliefs are then assumed to be consistent with this assumption ( $i$  believes  $j$  believes his opponent’s levels follow this distribution, *et cetera*). Strzalecki (2009) builds a similar—though more general—type-space model that encompasses all Level- $k$  models. It does not explicitly allow for levels to vary by game or for agents to update their beliefs upon observing signals, though both features could easily be incorporated.

$u_j(\sigma_i, \sigma_j)$  represent their expected payoffs. In some cases players receive informative signals about the type of their opponent; we represent  $i$ 's signal by  $\tau_i \in T$  and let  $\tau^0 \in T$  represent the uninformative 'null' signal.

Player  $i$ 's *type* is given by  $\theta_i = (c_i, k_i)$  where  $c_i : \Gamma \rightarrow \mathbb{N}_0 := \{0, 1, 2, \dots\}$  identifies  $i$ 's capacity for each game  $\gamma \in \Gamma$ , and  $k_i : \Gamma \times T \rightarrow \mathbb{N}_0$  identifies  $i$ 's level for each game  $\gamma \in \Gamma$  and signal  $\tau_i \in T$ . The capacity bounds the level, so  $k_i(\gamma, \tau_i) \leq c_i(\gamma)$  for all  $i$ ,  $\gamma$ , and  $\tau_i$ .<sup>5</sup> Let  $\Theta$  be the space of all possible types. Note that  $c_i$  does not vary in  $\tau_i$  since the capacity represents a player's underlying ability to 'solve' a particular game, regardless of the type of her opponent. The realized level  $k_i$  may vary in  $\tau_i$ , however, because the level stems directly from  $i$ 's belief about her opponent's strategy.

Beliefs are fixed by the model. Each player  $i$ 's pre-defined first-order beliefs are given by a mapping  $v : \mathbb{N}_0 \rightarrow \Delta(\mathbb{N}_0)$  such that  $v(k_i)(\{0, 1, \dots, k_i - 1\}) = 1$  for all  $k_i \in \mathbb{N}_0$ . For example, in Camerer et al. (2004),  $\lambda > 0$  is a free parameter and  $v(k)(l) = (\lambda^l / l!) / \sum_{\kappa=0}^{k-1} (\lambda^\kappa / \kappa!)$  if  $l < k$  and  $v(k)(l) = 0$  otherwise. The function  $v$  is common knowledge and therefore is not included in the description of  $\theta_i$ . Note that the  $k_i$  component of a player's type identifies which belief they have since changes in  $k_i$  lead to changes in  $v$ .

Behavior in a Level- $k$  model is defined inductively. The Level-0 strategy for each player  $i$  in  $\gamma$  is given exogenously as  $\sigma_i^0 \in \Delta(S_i)$ . If  $k_i(\gamma, \tau_i) = 0$  then player  $i$  plays  $\sigma_i^0$ . For each level  $k > 0$  the Level- $k$  strategy  $\sigma_i^k \in \Delta(S_i)$  for player  $i$  with  $k_i(\gamma, \tau_i) = k$  is the best response to beliefs  $v(k)$  given that each level  $\kappa < k$  of player  $j$  plays  $\sigma_j^\kappa$ .<sup>6</sup> Formally,  $\sigma_i^k$  for each  $k > 0$  is such that for all  $s'_i \in S_i$ ,

$$\sum_{\kappa=0}^{k-1} u_i(\sigma_i^k, \sigma_j^\kappa) v(k)(\kappa) \geq \sum_{\kappa=0}^{k-1} u_i(s'_i, \sigma_j^\kappa) v(k)(\kappa).$$

To see how this construction operates, fix a game  $\gamma$  and signal  $\tau_i$ . If player  $i$ 's type in this situation is  $(c_i, k_i) = (0, 0)$  then she plays  $\sigma_i^0$ . If  $i$ 's capacity is one then her type is either  $(1, 0)$  or  $(1, 1)$ . In the former case she plays  $\sigma_i^0$ ; in the latter case her beliefs are  $v(1)$ , which has  $v(1)(0) = 1$ , and so she plays  $\sigma_i^1$ . If  $i$ 's type is  $(2, 2)$  then she has beliefs  $v(2)$ , which puts pre-defined probabilities on her opponent being Level-0 and Level-1. In this case she plays  $\sigma_i^2$ . For any  $(c_i, k_i)$  player  $i$ 's beliefs are  $v(k_i)$  and her best response to those beliefs is  $\sigma_i^{k_i}$ . Note that beliefs depend only on  $k_i$ , so player types  $(4, 2)$ ,  $(3, 2)$ , and  $(2, 2)$  all have the same hierarchy of beliefs, for example.

<sup>5</sup>Technically, the inclusion of capacities is extraneous. A player's type could simply be defined as  $k_i : \Gamma \times T \rightarrow \mathbb{N}_0$  and then a capacity would then be derived by setting  $c_i(\gamma) = \sup_T k_i(\gamma, \tau_i)$  for each  $\gamma$ . We include capacities in the model to emphasize that agents' upper bounds on  $k_i$  may vary in  $\gamma$ .

<sup>6</sup>If there are multiple pure-strategy best responses then  $\sigma_i^k$  selects a distribution over those best responses, and that distribution is assumed to be common knowledge.

Once  $\sigma_i^0$  and  $v$  are defined, the only testable prediction of this model is that in each game and for each signal all players must select a strategy from the set  $\{\sigma_i^0, \sigma_i^1, \sigma_i^2, \dots\}$ .<sup>7</sup> In most applications, the researcher assumes that each level  $k$  plays  $\sigma_i^k$  with noise (usually with a logistic distribution) and then assigns each subject to the level that maximizes the sum of the likelihood across all games played.

Our goal is to consider a set of reasonable cross-game or cross-signal testable restrictions and explore which, if any, of these restrictions receive empirical support. All such restrictions can be expressed as properties of the functions  $k_i$  in the model. Examples of possible restrictions we can test using our experiments are:

- (1) **Constant:**  $k_i(\gamma, \tau_i) = k_i(\gamma', \tau'_i)$  for all  $i, \gamma, \gamma', \tau_i$ , and  $\tau'_i$ .
- (2) **Constant Across Games:**  $k_i(\gamma, \tau_i) = k_i(\gamma', \tau_i)$  for all  $i, \gamma, \gamma'$ , and  $\tau_i$ .
- (3) **Constant Ordering:** If  $k_i(\gamma, \tau) \geq k_j(\gamma, \tau)$  for some  $\gamma$  and  $\tau$  then  $k_i(\gamma', \tau') \geq k_j(\gamma', \tau')$  for all  $\gamma'$  and  $\tau'$ .
- (4) **Responsiveness to Signals:** For every  $\gamma$  and  $i$  there is some  $\tau$  and  $\tau'$  such that  $k_i(\gamma, \tau) > k_i(\gamma, \tau')$ .
- (5) **Consistent Ordering of Games:** For any  $\tau$ , if  $k_i(\gamma, \tau) \geq k_i(\gamma', \tau)$  for some  $i, \gamma$  and  $\gamma'$ , then  $k_j(\gamma, \tau) \geq k_j(\gamma', \tau)$  for all  $j$ .

The first restriction represents a very strict interpretation of the Level- $k$  model in which each person's level never varies, regardless of the difficulty of the game or the information received. The second restriction weakens the first by allowing players' beliefs to respond to information.

Instead of forcing absolute levels to be constant, the third restriction requires only that players' relative levels be fixed. Thus, if Anne plays a (weakly) higher level than Bob in one game when they have identical information then Anne should play a (weakly) higher level than Bob in all games where they have identical information. Certainly this would be violated with differing degrees of game-specific experience; recall, however, that the Level- $k$  model applies only to the first-time play of novel games.<sup>8</sup>

The fourth restriction requires that there exist a pair of signals in each game over which a player's level will differ. Thus, a minimal amount of responsiveness to information is assumed.

<sup>7</sup>If  $\sigma^0$  is not restricted then there are no testable predictions; letting  $\sigma^0$  equal the empirical distribution of strategies provides a perfect fit.

<sup>8</sup>Cross-game learning may still generate violators of this restriction; a chess master may play to a higher level than a professional soccer player in checkers, but to a lower level in an asymmetric matching pennies game. For this reason the boundaries of applicability of the Level- $k$  model are sometimes ambiguous.

	1	2	3	4	5	6	7
1	1	10	0	0	0	0	-11
	1	-10	0	0	0	0	0
2	-10	0	10	0	0	0	0
	10	0	-10	0	0	0	0
3	0	-10	0	10	0	0	0
	0	10	0	-10	0	0	0
4	0	0	-10	0	10	10	10
	0	0	10	0	-10	-10	-10
5	0	0	0	-10	0	0	0
	0	0	0	10	0	0	0
6	0	0	0	-10	0	0	0
	0	0	0	10	0	0	0
7	0	0	0	-10	0	0	-11
	-11	0	0	10	0	0	-11

FIGURE I. Undercutting game 1 (UG1).

The last restriction listed implies that the observed levels can be used to order the games in  $\Gamma$ . If, at some fixed signal, all players play a lower level in  $\gamma'$  than in  $\gamma$  then it can be inferred that  $\gamma'$  is a more difficult or complex game.

It is certainly easy to imagine plausible variations on the function  $k_i$  that violate each of these restrictions, or that violate any other restriction we may consider. But each restriction that is violated means the loss of a testable implication for the model. If the most empirically accurate version of the Level- $k$  model requires  $k_i$  functions that satisfy no cross-game or cross-signal restrictions then the model cannot be used to make out-of-sample predictions about behavior. Thus, the predictive power of the model hinges on the presence of at least some such restrictions.

#### IV THE GAMES

We study two families of games: a novel family of games that are useful for identifying player types, which we call undercutting games (UG), and the two-person guessing games (2PGG) studied by Costa-Gomes and Crawford (2006).

##### *Undercutting Games*

An undercutting game is a symmetric, two-player game parameterized by two positive integers  $m$  and  $n$  with  $m < n$ . Each player  $i \in \{1, 2\}$  picks a positive integer  $s_i \in \{1, 2, \dots, m, \dots, n\}$ . Player  $i$  wins \$10 from player  $j$  if either  $s_i = m < s_j$  or  $s_i + 1 = s_j \leq m$ . Thus, if player  $i$  expects her opponent to choose  $s_j > m$  then she best responds by choosing  $s_i = m$ ; otherwise she best responds by ‘undercutting’ her opponent and choosing  $s_i = s_j - 1$ . If no player ‘wins’ then one of the following situations apply: If both choose

	1	2	3	4	5	6	7	8	9
1	1	10	0	0	0	0	0	0	-11
2	1	-10	0	0	0	0	0	0	0
3	0	0	10	0	0	0	0	0	0
4	0	0	0	10	0	0	0	0	0
5	0	0	0	0	10	0	0	0	0
6	0	0	0	0	0	10	0	0	0
7	0	0	0	0	0	0	10	0	0
8	0	0	0	0	0	0	0	10	0
9	0	0	0	0	0	0	0	0	10

FIGURE II. Undercutting game 2 (UG2).

	1	2	3	4	5	6	7	8	9
1	1	10	0	0	0	0	0	0	-11
2	1	-10	0	0	0	0	0	0	0
3	0	0	10	0	0	0	0	0	0
4	0	0	0	10	0	0	0	0	0
5	0	0	0	0	10	0	0	0	0
6	0	0	0	0	0	10	0	0	0
7	0	0	0	0	0	0	10	0	0
8	0	0	0	0	0	0	0	10	0
9	0	0	0	0	0	0	0	0	10

FIGURE III. Undercutting game 3 (UG3).

	1	2	3	4	5	6	7
1	1	10	0	0	0	0	-11
2	1	-10	0	0	0	0	0
3	0	0	10	0	0	0	0
4	0	0	0	10	0	0	0
5	0	0	0	0	10	0	0
6	0	0	0	0	0	10	0
7	0	0	0	0	0	0	10

FIGURE IV. Undercutting game 4 (UG4).

one (the unique Nash equilibrium choice) then both earn a payoff of one. If both choose

$n$  then both lose 11. If  $i$  chooses one and  $j$  chooses  $n$  then  $i$  loses 11 and  $j$  earns nothing. In all other cases both players earn zero. These cases are designed to rule out any mixed-strategy Nash equilibria, making  $(1, 1)$  the unique Nash equilibrium of the game.

The payoff matrices of the undercutting games used in this experiment are shown in Figures I–IV. Consider UG1, shown in Figure I. A levels-of-reasoning model that assumes uniformly random play by Level-0 types will predict that Level-1 types play  $m = 4$  as it maximizes the sum of row payoffs, Level-2 types play 3, Level-3 types play 2, and all higher levels play the equilibrium strategy of 1. This enables a unique identification of a player’s level (up to Level 4, except for UG3 where levels 5 and 6 are also possible) from a single observation of their strategy.

The game in Figure IV, UG4, departs from UG2 only in that three dominated actions have been ‘compressed’ into one (which is now itself also dominated by another dominated action). Since dominated actions are never predicted for types above Level-0, this modification should have little impact on the distribution of types.

This family of games was designed explicitly for testing the Level- $k$  model. Its undercutting structure is intended to focus players’ attention on the strategies of their opponents, encouraging Level- $k$ -type thinking. The strategy space is relatively small, unlike p-beauty contest games, but the only strategy that confounds multiple levels (other than the Level-0 type, which may randomize over many strategies) is the Nash equilibrium strategy since all levels greater than  $m$  are predicted to play this action. There are no other Nash equilibria in pure or mixed strategies. And variations in the Level-0 strategy simply shift all levels uniformly; different Level-0 models may assign different levels to different players, but will not alter the relative ordering of players’ levels.

### *Two-Person Guessing Games*

Two-person guessing games are asymmetric, two-player games parameterized by a lower bound  $a_i \geq 0$ , upper bound  $b_i > a_i$ , and target  $p_i > 0$  for each player. Strategies are given by  $s_i \in [a_i, b_i]$  and player  $i$  is paid according to how far her choice is from  $p_i$  times  $s_j$ , denoted by  $e_i = |s_i - p_i s_j|$ . We study two versions of these guessing games: the standard two-person guessing game using the payoff function from Costa-Gomes and Crawford (2006) and a novel zero-sum version of the two-person guessing game.

In the standard version each player  $i$ ’s payment is a quasiconcave function of  $e_i$  that is maximized at zero. Specifically, players receive  $15 - (11/200)e_i$  dollars if  $e_i \leq 200$ ,  $5 - (1/200)e_i$  dollars if  $e_i \in (200, 1000]$ , and zero if  $e_i \geq 1000$ . The unique best response is to set  $e_i = 0$  by choosing  $s_i = p_i s_j$ . If  $p_i s_j$  lands outside of  $i$ ’s strategy space then

the nearest endpoint of the strategy space is the best response. In a levels-of-reasoning model, Level-0 may be assumed to randomize uniformly over  $[a_i, b_i]$  or to play the midpoint of  $[a_i, b_i]$  with certainty. In either case Level-1 types will play  $p_i$  times  $(a_j + b_j)/2$ ; if this is not attainable then the Level-1 player will select the nearest endpoint of her interval. A Level-2 type will play  $p_i p_j$  times her own midpoint (or the nearest endpoint), and so on. This iterative reasoning converges to a Nash equilibrium with one player playing on the boundary of her interval and the other best-responding to that boundary strategy (see Costa-Gomes and Crawford, 2006).

In the zero-sum version both  $e_i$  and  $e_j$  are calculated and if  $e_i > e_j$  then  $i$  pays \$10 to  $j$ ; if  $e_i < e_j$  then  $j$  pays \$10 to  $i$ . If  $e_i = e_j$  then a winner is randomly chosen who receives \$10 from the loser. In this game it is still a best-response to play  $s_i = p_i s_j$ , but this best response is far from unique since any  $s_i$  that guarantees  $e_i < e_j$  will also be a best response. Intuitively, a strategy that gives a large value of  $e_i$  may be acceptable as long as  $e_j$  is larger. Under some parameterizations there are even dominant strategies  $s_i$  that guarantee  $e_j > e_i$  regardless of  $s_j$ .<sup>9</sup>

## V EXPERIMENTAL DESIGN

In total we test 174 subjects, all undergraduates from a variety of majors at The Ohio State University. At the beginning of the experiment (after instructions are read aloud), each subject takes five quizzes:

- (1) an IQ test to measure general cognitive ability,
- (2) the Eye Gaze test for adult autism,
- (3) the Wechsler digit span working memory test,
- (4) the Cognitive Reflection Test (CRT), and
- (5) the one-player Takeover game.

Each of these quizzes represents a previously-used measure of general intelligence or strategic sophistication. The IQ test consists of ten questions taken from the Mensa society's 'workout' exam.<sup>10</sup> The Eye Gaze test (Baron-Cohen et al., 1997) asks subjects to identify the emotions being expressed by a pair of eyes in a photograph. See Figure V for sample problems. Poor performance on this task is diagnostic of high-functioning adult

<sup>9</sup>As a simple example, consider a zero-sum guessing game with  $a_1 = 0$ ,  $p_1 < 1$ , and  $a_2 > 0$ . If player 1 chooses  $s_1 = 0$  then  $e_1 = p_1 s_2$  and  $e_2 = s_2$ , so player 1 is guaranteed to win. In fact, player 1 is guaranteed to win for any  $s_1 \leq a_2$ . Note, however, that a Level-1 player 1 will choose  $s_1 = p_1(a_2 + b_2)/2$  with positive probability, even if that choice is weakly dominated by  $s_1 = 0$ . This variation in payoffs allow us to test the sensitivity of behavior to the best-response structure of the two-person guessing games.

<sup>10</sup>See <http://www.mensa.org/workout2.php>

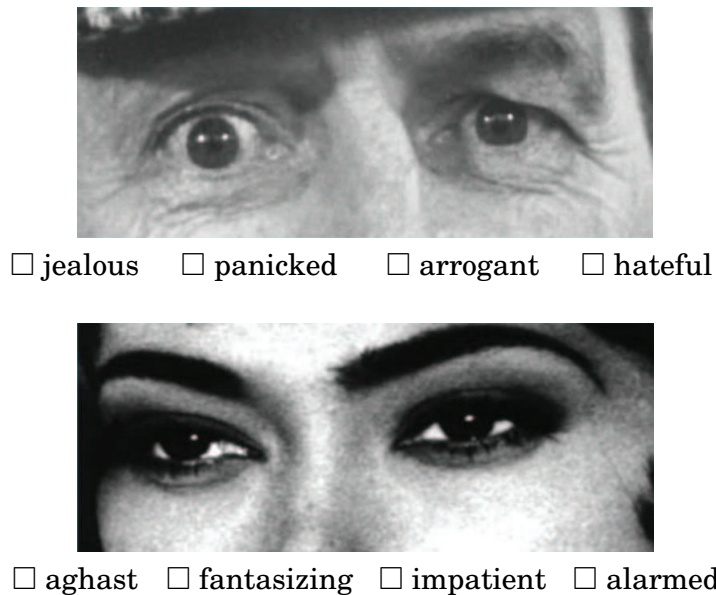


FIGURE V. Sample questions from the Eye Gaze test.

autism or Asperger's Syndrome (Baron-Cohen et al., 1997) and strong performance is correlated with the ability to determine whether or not price movements in a market are affected by a trader with inside information (Bruguier et al., 2008).

The Wechsler digit span memory test tests subjects' abilities to recall strings of digits of increasing length; this task has been used as one measure of human intelligence (Wechsler, 1958). Devetag and Warglien (2003) had 67 subjects take this short-term memory test and then play three games against a computerized opponent that always selects the equilibrium strategy. The three games all require iterated reasoning to solve the equilibrium best response. The correlation between subjects' memory test score and the frequency with which they select the best response is positive (Kendall's  $\tau$ : 0.248) and significant ( $p$ -value: 0.010).

The CRT contains three questions for which the 'knee-jerk' response is often wrong. Performance on the test is correlated with measured time preferences, risk taking in gains, risk aversion in losses, and other IQ measures (Frederick, 2005). This measure also correlates with a tendency to play default strategies in public goods games (Altmann and Falk, 2009).

Finally, the one-player Takeover game is a single-player adverse selection problem in which the subject is asked to make an offer to buy a company knowing that the seller will only sell if the company's value is less than the offer. Given the parameters of the problem, all positive offers are unprofitable in expectation, yet many subjects fall victim

Game ID	Game Type	Player's Limits & Target	Opponent's Limits & Target
UG1	Undercutting Game	See Figure I	
UG2	Undercutting Game	See Figure II	
UG3	Undercutting Game	See Figure III	
UG4	Undercutting Game	See Figure IV	
GG5	Guessing Game	([215, 815], 1.4)	([0, 650], 0.9)
GG6	Guessing Game	([100, 500], 0.7)	([300, 900], 1.3)
GG7	Guessing Game	([100, 500], 0.5)	([100, 900], 1.3)
GG8	Guessing Game	([0, 650], 0.9)	([215, 815], 1.4)
GG9	Guessing Game	([300, 900], 1.3)	([100, 500], 0.7)
GG10	Guessing Game	([100, 900], 1.3)	([100, 500], 0.5)

TABLE I. The ten games used in the experiment.

to the ‘winner’s curse’ by submitting positive offers (Samuelson and Bazerman, 1985), even after receiving feedback and gaining experience (Ball et al., 1991).

Each of the quiz scores is normalized to a scale of ten possible points. For scoring purposes during the experiment, the CRT and Takeover game were combined into one four-question, ten-point quiz since the one question in the Takeover game quiz would receive disproportionate weight were it scored separately out of ten points. In the data analysis below we disaggregate these two quizzes. In our analysis we use a score for the Takeover game that is linearly decreasing in a subject’s bid; in the experiment subjects received a positive score for this question if and only if their bid was exactly zero—the unique profit-maximizing bid.<sup>11</sup> The sum of the four quiz scores was calculated for each player. Players were given no feedback about any player’s absolute or relative performance on the quizzes until the end of the experiment, at which point they learned only their own total quiz score.

After completing the quizzes subjects then played ten games. The list of games is shown in Table I. Note that although the guessing games are asymmetric, players actually played both roles of each game, as in Costa-Gomes and Crawford, 2006. Players’ decisions in GG5, for example, were matched with another player’s player-1 decision in GG8 to determine payoffs. Sixty-eight of the subjects faced the standard version of the six two-person guessing games while 106 subjects faced the zero-sum version of the same six guessing games.

<sup>11</sup>Specifically, a subject who submitted a bid of  $b_i$  was scored as earning  $10(1 - b_i/\max_j b_j)$  points in our analysis.

With zero-sum payoffs, GG5 has dominant strategies for player 2: any  $s_2 \in \{0, \dots, 53\} \cup \{646, \dots, 650\}$  guarantees a victory regardless of  $s_1$ . GG6 has dominant strategies for player 1: any  $s_1 \in \{100, \dots, 221\}$  guarantees a victory regardless of  $s_2$ . GG7 has no dominant strategy for either player. Recall that GG8, GG9, and GG10 are simply the opposite-role versions of GG5, GG6, and GG7, respectively. Thus, among those playing the zero-sum variants, each subject plays two guessing games that have dominant strategies for themselves, two in which their opponent has dominant strategies, and two in which neither player has a dominant strategy.

The Level- $k$  model does not predict singleton strategies with zero-sum payoffs since a large set of strategies guarantees a ‘win’ against a certain belief about one’s opponent. Assuming players randomize uniformly over all best responses, the strategies do not converge to equilibrium strategies as  $k$  is increased for our games. Instead, strategies oscillate in  $k$ , with all even levels playing one strategy and all odd levels playing another.<sup>12</sup> Given the non-identifiability of individual levels from single actions in these zero-sum games, we omit these data when estimating levels unless otherwise stated. Our analysis of these data focuses instead on a comparison of behavior with the standard guessing games, and on whether players identify the dominant strategies.

In each game subjects were asked to choose a strategy against a random opponent, against the opponent (other than themselves) with the highest total score on all of the quizzes, and the opponent (other than themselves) with the lowest score on the quizzes. All choices were made without feedback. After making these three choices in all 10 games, players learned that they could ‘loop back’ through the games and revise their choices if desired. This could be done up to four times, for a total of five iterations through the 10 games, each without feedback. Once subjects finished all five iterations or declined the opportunity to loop back, their play was recorded, 4 of their choices were randomly selected (two from the undercutting games and two from the guessing games), and they were matched with another player and paid for their decisions. Subjects earning less than \$6 (the standard show-up fee) were paid \$6 for their time. Subjects earned an average of \$26.44 in total.

Instructions and screen shots from each of the quizzes and games are available in an accompanying appendix.

---

<sup>12</sup>In the games with dominant strategies the player with a set of dominant strategies plays exactly those strategies for all odd levels, but also plays some weakly dominated best responses for all even levels. The player without a dominant strategy randomizes uniformly in all even levels (all strategies are equally bad when players believe their opponent will win for sure) and plays a unique best response in the odd levels where he believes his opponent to be playing some weakly dominated strategies as well.

## VI DATA ANALYSIS PROCEDURES

Our experiment consists of sixteen games ( $\Gamma = \{UG1, \dots, UG4, GG5, \dots, GG10, ZSGG5, \dots, ZSGG10\}$ ), of which each subject sees ten. We employ three possible signals ( $T = \{\tau^{LO}, \tau^0, \tau^{HI}\}$ ) indicating whether the current opponent has the lowest quiz score, is randomly selected, or has the highest quiz score. To each  $i \in \{1, \dots, 174\}$ ,  $\gamma \in \Gamma$  and  $\tau \in T$  we assign a level  $k_i(\gamma, \tau)$  using a simple maximum-likelihood approach that follows closely CGC06. As in CGC06 (and several other papers), we focus on the case where  $v(k)(k-1) = 1$  for all  $k > 0$  and  $\sigma_i^0$  is uniform over  $S_i$ . For the undercutting games the identification of levels is fairly robust to these assumptions. The robustness of the levels in the guessing game to these specifications is not explored here. For simplicity, we aggregate all observed levels four and above into a category we refer to as ‘Level-4+’.

For undercutting games the maximum likelihood approach identifies levels with actions directly. Anyone playing a dominated strategy is labeled a Level-0 and all others are labeled according to the strategy they chose. This skews the population frequencies away from Level-0 since Level-0 types also play undominated strategies with positive probabilities, but for each individual data point this method identifies the maximum likelihood level  $k_i(\gamma, \tau)$ .

For the guessing games we follow CGC06’s econometric specification in our assignment of levels, differing only in that we identify a level for each person in each game, while CGC06 estimates one level per person across all games, and that we allow for Level-0 types while CGC06 does not. For each player, each level  $k$  is modeled as playing  $\sigma_i^k$  with probability  $1 - \varepsilon$  and playing a random strategy with probability  $\varepsilon$ . The probability distribution over random strategies is given by the logistic distribution so that actions with a higher expected payoff against the Level- $k$  beliefs  $v(k)$  are played with a higher probability. For each level we calculate the values of  $\varepsilon$  and the free parameter in the logistic distribution ( $\lambda$ ) that maximize the likelihood score for that level. The level with the highest likelihood value is then assigned to that subject for that game. If the likelihood value is less than the likelihood value assuming a uniform distribution over the strategy space, or if the maximum likelihood value of the logistic distribution parameter is 0.01 (the minimum allowable in our grid search), then the subject is labeled as a Level-0 for that game. See CGC06 for details.

It is important to note that our procedure for assigning a level to each observation is not an econometric estimation, but rather an assignment rule based on the underlying econometric model of CGC06. The assigned values of  $k$ ,  $\varepsilon$ , and  $\lambda$  have no econometric

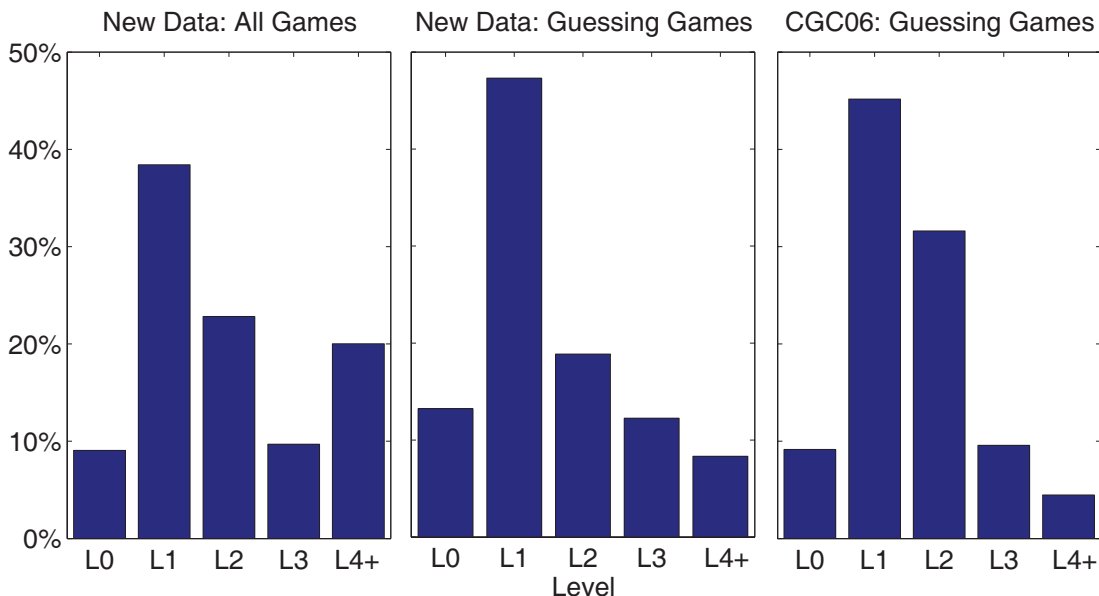


FIGURE VI. Aggregate level distributions across all ten games, across only the six guessing games, and for the Costa-Gomes and Crawford (2006) data.

interpretation since, as an econometric model, ours would have three estimated parameters per observation and therefore would be grossly over-identified. As a robustness check, we estimated levels in the standard guessing games fixing  $\varepsilon = 1$ , setting  $\lambda$  to 1.33 (the average estimated value of  $\lambda$  in CGC06 using only subjects' guesses), and then estimating  $k$  for each observation. Under this new procedure, 85.5% of observations received the same level assignment as in our original procedure. Roughly half of the observations whose level did change became Level-0 observations, implying their likelihood value simply fell below the uniform distribution likelihood. None of the key results of the paper changed under these alternative estimates.

## VII RESULTS

### *Aggregate Distributions of Levels*

Each subject in each game can be identified with a maximum-likelihood level based on the action the subject took in that one game. Level distributions aggregated across

Game	L0	L1	L2	L3	L4+/Nash
UG1	7.47%	35.06%	21.84%	9.77%	25.86%
UG2	6.90%	36.78%	25.29%	7.47%	23.56%
UG3	4.02%	28.16%	25.29%	8.05%	34.48%
UG4	8.05%	32.76%	28.16%	7.47%	23.56%
Total UG	6.61%	33.19%	25.14%	8.19%	26.87%
GG5*	5.88%	73.53%	8.82%	8.82%	2.94%
GG6*	1.47%	64.71%	13.24%	14.71%	5.88%
GG7*	51.47%	30.88%	14.71%	1.47%	1.47%
GG8*	7.35%	39.71%	25.00%	20.59%	7.35%
GG9*	7.35%	39.71%	22.06%	4.41%	26.47%
GG10*	5.88%	35.29%	29.41%	23.53%	5.88%
Total GG*	13.24%	47.30%	18.87%	12.25%	8.33%
Total	10.58%	41.66%	21.38%	10.63%	15.75%

\*Excludes zero-sum guessing game data.

TABLE II. Frequency of levels in each game.

games are shown in Figure VI alongside the distribution from CGC06.<sup>13</sup> The distributions for the new guessing game data exhibit the typical pattern: Level 1 is the modal type with frequencies descending for higher types. Qualitatively, the distributions are similar to those of CGC06.<sup>14</sup>

However, the aggregate histogram hides a significant amount of variance between the individual games, as shown in Table II. In particular, for the guessing games the distribution is clearly not stable. For example, the fraction of Level-0 play jumps from 1.47% in guessing game 6 (GG6) to 51.47% in GG7. The fraction of Level-1 play more than doubles from 30.88% in GG7 to 73.53% in GG5. Nash play in GG5 is 2.94% but rises to 26.47% in GG9. The distribution appears more stable between undercutting games, where levels are exactly inferred from a small number of available actions rather than from a continuous strategy space.

The cross-game variance of distributions is not unique to our data; per-game estimated type distributions from CGC06's data are shown in Table III. In games 2–8 we see no Nash equilibrium types, while in game 13 we see over 20% Nash types. Level 1 play varies from 21.59% up to 73.86%.

<sup>13</sup>The histogram of CGC06 data is generated using our procedure of identifying one type per person per game (allowing for Level-0 types) and then generating a histogram of all player-game observations; CGC06 instead estimate only one type per person (excluding Level 0) based on their play in all games.

<sup>14</sup>The quantal response equilibrium (McKelvey and Palfrey, 1995) also fits our aggregate data reasonably well; we do not explore cross-game stability of that model in this paper.

Game	L0	L1	L2	L3	Nash
1	7.95%	47.73%	12.50%	19.32%	12.50%
2	14.77%	21.59%	45.45%	18.18%	0.00%
3	14.77%	55.68%	18.18%	11.36%	0.00%
4	14.77%	35.23%	50.00%	0.00%	0.00%
5	14.77%	73.86%	4.55%	6.82%	0.00%
6	7.95%	54.55%	37.50%	0.00%	0.00%
7	9.09%	62.50%	26.14%	2.27%	0.00%
8	5.68%	71.59%	20.45%	2.27%	0.00%
9	13.64%	38.64%	40.91%	2.27%	4.55%
10	0.00%	37.50%	32.95%	26.14%	3.41%
11	10.23%	36.36%	46.59%	2.27%	4.55%
12	1.14%	45.45%	34.09%	18.18%	1.14%
13	4.55%	23.86%	40.91%	10.23%	20.45%
14	10.23%	35.23%	28.41%	18.18%	7.95%
15	7.95%	36.36%	30.68%	13.64%	11.36%
16	9.09%	46.59%	36.36%	2.27%	5.68%
Total	9.16%	45.17%	31.61%	9.59%	4.47%

TABLE III. Frequency of estimated levels in each game of Costa-Gomes and Crawford (2006).

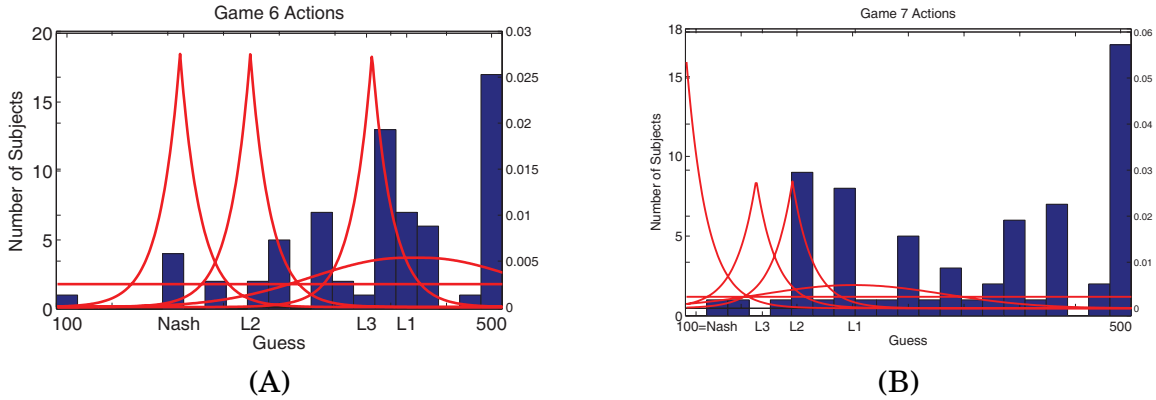


FIGURE VII. Histograms of actions in (A) GG6 and (B) GG7 along with logistic response functions for each level assuming  $\lambda = 1$ .

Figure VII shows how types are assigned in the guessing games using GG6 and GG7 as examples. In each game the solid bars represent the histogram of actual actions over player 1's interval of  $[0, 500]$ . The flat line represents the uniform distribution of Level-0 actions. Assuming a logistic response with error parameter  $\lambda = 1$  gives a Level-1 action distribution which is quite flat; it is centered above the label 'L1' in the graphs. The Level-2, Level-3, and Nash action distributions (with  $\lambda = 1$ ) have much lower variance

Data	L1	L2	L3	L4+	Total
New Data	5.88%	2.45%	0%	5.64%	3.49%
CGC06	25.36%	21.16%	13.71%	17.61%	19.46%

TABLE IV. Frequency of exact conformity with the Level- $k$  model in the new data and in Costa-Gomes and Crawford (2006).

because those player types have point predictions about their opponents' strategy, making mistakes from the best response more costly. The Level- $k$  model appears to perform quite poorly in GG7; most data are categorized as Level-0 or Level-1 simply because those levels' predicted actions have higher variance.

The histograms in Figure VII reveal that the modal play is 500 in both games. This may be justified by a model in which confused players mistakenly believe their target to equal 1.0, leading players to guess their opponent's strategy directly rather than some multiple of that strategy. In other games, however, this explanation does not organize the data; only 2 subjects behave this way in GG5 and GG8 and no subject acts as a Level-1 type using a target of 1.0 in more than 3 of the 6 guessing games.

Many players select guesses that are even multiples of 100. Out of 408 observations this occurs 42.4% of the time. In contrast, an even multiple of 100 is predicted by the Level- $k$  model only for the Nash type and only in games 7 and 9. Using our population frequency of Nash types, this implies that even multiples of 100 should be observed in roughly 5% of the observations. Costa-Gomes and Crawford (2006) also observe a high frequency of multiples-of-100 play, in 35.9% of their observations, though it is predicted for their games 22.7% of the time (using their estimates of the type distribution).

In the undercutting games, the distribution of types is generally much more stable across games. In all four games, there is a high proportion of L1, L2 and L4+/Nash behavior, and very little behavior corresponding to L0. In these games, actions uniquely identify types, but all dominated actions are associated with the Level-0 type. This under-counts the Level-0 type since Level-0 players should also randomize over non-dominated strategies, but the maximum likelihood type for a non-dominant strategy is never Level 0. Thus, the frequencies of Level-0 types could be adjusted by dividing by the fraction of strategies which are dominated. Adjusted Level-0 frequencies for the four undercutting games are 17.43%, 12.42%, 12.06%, and 14.49%, respectively.

*Exact Hits*

Costa-Gomes and Crawford (2006) find that roughly one-fifth of their subjects play exactly as predicted by the Level- $k$  model. Table IV reveals that this is not true for the current data; less than 4% of our guessing game data exactly conformed to a Level- $k$  prediction. Although this may represent a subject pool difference (San Diego and York versus Ohio State) we conjecture that the degree of training given in the instructions is driving this effect. CGC06's subjects read through 19 pages of instructions that included questions on how to calculate best responses to particular strategies of opponents. Our instructions were 5 pages long and only informed subjects of how their payoffs are calculated; we did not explicitly train subjects to calculate best responses. If this is true then the accuracy of the Level- $k$  model depends on the level of training agents receive prior to choosing their strategies. In many situations this may not be observable, making predictions of strategies difficult.

*Dominated Strategies*

While we exclude the zero-sum guessing game data from the type classification above, this data allows us to explore the frequency with which subjects play dominated strategies and the extent to which their behavior is sensitive to the varying best-response characteristics of the guessing games. In the zero-sum guessing games in which there is a dominant strategy (ZSGG6 and ZSGG8) only 10.9% of choices are dominant strategies (11.3% and 10.4%, respectively). Only 4 of the 106 subjects (3.8%) played a dominant strategy in both games. Thus, only a very small proportion of subjects made choices consistent with any level higher than L0 when we changed the best-response structure of the game.

In fact, the distribution of guesses is roughly the same between those playing the standard guessing game and those playing the zero-sum guessing game; a Kolmogorov-Smirnov test for differences gives  $p$ -values of 0.70 and 0.24 for games 6 and 8, respectively. Games 5 and 9 also have insignificant differences, though the K-S test does reveal significant differences for games 7 and 10 ( $p$ -values of 0.005 and 0.014, respectively). (Recall that game 10 is simply game 7 with the subjects playing the role of player 2.) Thus, dramatic changes in the best response function can have little to no effect on chosen actions in certain games, implying that any behavioral model that depends solely on best-response behavior will be unable to explain these observations.

From ↓ To →	L0	L1	L2	L3	L4+
L0	<b>40.6%</b>	23.9%	9.4%	10.9%	15.2%
L1	4.8%	<b>62.3%</b>	15.7%	4.6%	12.6%
L2	2.5%	20.8%	<b>60.6%</b>	8.8%	7.4%
L3	8.8%	18.7%	<b>26.9%</b>	22.2%	23.4%
L4+	3.7%	15.5%	7.0%	7.1%	<b>66.7%</b>

TABLE V. Markov transition between levels for the four undercutting games.

From ↓ To →	L0	L1	L2	L3	L4+
L0	8.9%	<b>48.1%</b>	17.0%	12.6%	13.3%
L1	13.5%	<b>50.2%</b>	18.2%	11.4%	<b>6.7%</b>
L2	11.9%	<b>45.7%</b>	24.4%	10.6%	7.3%
L3	13.6%	<b>44.0%</b>	16.4%	14.4%	11.6%
L4+	21.2%	<b>38.2%</b>	16.5%	17.1%	7.1%
Any (if i.i.d.)	13.2%	<b>47.3%</b>	18.9%	12.3%	8.3%

TABLE VI. Markov transition between levels for the six standard two-person guessing games.

### *Persistence of Absolute Levels*

At the individual level we can look at how frequently players switch levels between games, violating the testable restriction that  $k_i(\gamma, \tau^0) = k_i(\gamma', \tau^0)$  for all  $\gamma$  and  $\gamma'$ . We generate a Markov transition matrix by selecting all pairs of games (played against random opponents, so  $\tau_i = \tau^0$ ) and for each level of one game calculating the frequency with which a subject moves to each level in the other game. Tables V and VI show these transition matrices for the undercutting and guessing games, respectively. Clearly, players' levels are more stable in the undercutting games than in the guessing games. In fact, Level 1 acts as an absorbing state in the guessing games. The last row of Table VI shows that this result is relatively consistent with independently-chosen random actions in each game, given the flatness of the Level-1 logistic response function (see Figure VII). Thus, we conclude that absolute levels are stable across some families of similar games, but not all;  $k_i$  is not constant in  $\gamma$ .

### *Persistence of Relative Levels*

To examine the frequency with which the ordinal ranking of players' levels changes between games, we randomly draw two games and two players and measure the frequency with which the strictly higher-level player in one game becomes the strictly lower-level

	Frequency	i.i.d. Prob.
Standard Guessing Games		
Same level in both games:	14.0%	8.9%
Same level in one game:	44.7%	41.9%
Higher level stays higher:	21.7%	24.6%
Higher level becomes lower:	19.7%	24.6%
Switch/non-switch ratio:	0.908	1.00
Undercutting Games		
Same level in both games:	11.1%	6.6%
Same level in one game:	29.0%	38.2%
Higher level stays higher:	46.5%	27.6%
Higher level becomes lower:	13.4%	27.6%
Switch/non-switch ratio:	0.288	1.00
Both Families of Games		
Same level in both games:	8.5%	5.3%
Same level in one game:	37.5%	35.6%
Higher level stays higher:	30.8%	29.6%
Higher level becomes lower:	23.2%	29.6%
Switch/non-switch ratio:	0.753	1.00

TABLE VII. Observed frequency of level-switching among pairs of subjects between randomly-drawn games, compared to the expected frequency under independently-drawn (i.i.d.) types.

player in another. Table VII shows the frequency with which the two players do not strictly switch levels (either by having the same level in both games or in one game), the frequency with which the strictly higher-level player in one game stays the strictly higher-level player in the other, and the frequency with which the players strictly switch the ordering of their levels. These frequencies are compared to the probabilities with which these events would occur if types were drawn independently from the observed population distribution of levels.

If  $k_i(\gamma, \tau^0) \geq k_j(\gamma, \tau^0)$  implies  $k_i(\gamma', \tau^0) \geq k_j(\gamma', \tau^0)$  for all  $\gamma'$  then no strict switching of levels should occur. If levels are independently drawn in each game then the probability of a strict switch in two players' levels exactly equals the probability that the two players have strictly different levels whose ordering does not change between games. In the guessing games, for example, persistent relative levels implies that the higher-level subject should never become the lower-level subject, whereas with randomly-drawn types this switch will occur in 24.6% of the pairs. In fact we observe a strict switch in 19.7% of the pairs. Furthermore, the ratio of strict switches to strict non-switches is 0.908, close to the one-to-one ratio predicted if levels were randomly drawn.

	Const.	IQ	EyeGaze	Memory	CRT	Takeover
Coefficient	-4.572	-0.362	<b>1.991</b>	0.528	<b>0.605</b>	<i>0.815</i>
p-value	(0.318)	(0.351)	<b>(0.005)</b>	(0.155)	<b>(0.035)</b>	<i>(0.069)</i>

TABLE VIII. Regression of total earnings on the five quiz scores.

Since absolute levels in the undercutting games are fairly stable, we expect similar persistence in subjects' relative levels. In fact, strict switches occur in 13.4% of the pairs and the ratio of strict switches to strict non-switches is 0.288, far from the 27.6% and one-to-one ratio predicted by a random-levels model.

Pooling both families of games obviously gives an intermediate result: A pair of subjects that don't play the same level in either game have a 23.2% chance of switching their order between games and a 30.8% chance of maintaining their order (compared to the 29.6% random-level benchmark), yielding a ratio of 0.753.

### *Using Quizzes to Predict Levels*

Each subject took five quizzes: an IQ quiz, the Eye Gaze quiz, a memory quiz, the Cognitive Reflection Test (CRT), and a one-player Takeover Game. Each quiz score is normalized to a ten-point scale, and a total quiz score is calculated by summing these normalized scores.

We first test for correlations between the various quizzes. We find that the memory quiz is positively correlated with both the IQ quiz and the CRT quiz, with Spearman rank correlations (and  $p$ -values) of 0.151 (0.047) and 0.170 (0.025), respectively, and that the CRT quiz is positively correlated with the Takeover Game score (rank correlation of 0.194 with a  $p$ -value of 0.010). No other correlations are significantly different from zero.

The total quiz score is positively correlated with subjects' earnings (rank correlation of 0.295 with a  $p$ -value of less than 0.001), and a linear regression (Table VIII) shows that earnings are most strongly correlated with the Eye Gaze and CRT quiz scores, with a weaker correlation for the Takeover game score.

Subjects' median levels are not significantly correlated with individual quiz scores. Thus, we reject the joint hypothesis that these quizzes measure strategic sophistication and that sophistication is monotonically related to the realized level of play. Looking

	Const.	IQ	EyeGaze	Memory	CRT	Takeover
Undercutting Games						
Coefficient	1.622	0.079	<b>-0.303</b>	<b>-0.167</b>	-0.074	0.025
p-value	(0.299)	(0.614)	<b>(0.013)</b>	<b>(0.031)</b>	(0.190)	(0.775)
Standard Guessing Games						
Coefficient	1.161	0.268	<i>-0.347</i>	-0.181	0.018	-0.086
p-value	(0.678)	(0.354)	<i>(0.080)</i>	(0.180)	(0.823)	(0.552)

TABLE IX. Logistic regressions of the probability of playing the Level-1 strategy in a majority of games as a function of observed quiz scores.

at earnings for each level, however, we find that playing high levels is not a payoff-maximizing strategy in any game. In most cases the Level-2 strategy is the empirical best response. Thus, a truly sophisticated player with a high level *capacity* should choose to play lower-level strategies to maximize their earnings.

Given that payoffs are not monotonic in levels, we look instead at whether quiz scores can predict individual levels using logistic regressions. First, for the two families of games, undercutting games and standard guessing games, we ask whether quiz scores have a significant impact on the probability that a player plays the Level-1 strategy in a strict majority of games in that family. In the undercutting games we see that players who score poorly on the Eye Gaze and memory quizzes are more likely to be Level-1 players, with the Eye Gaze correlation being larger in magnitude. In guessing games the results are weaker, though the Eye Gaze score is weakly negatively correlated with the propensity for frequent Level-1 play.

The Eye Gaze correlation with Level-1 play has intuitive appeal: Poor performance in the Eye Gaze quiz is diagnostic of adult autism (Baron-Cohen et al., 1997) and autism is often characterized by a lack of ‘Theory of Mind’ (Baron-Cohen, 1990), meaning autistic people fail to recognize that others behave in response to conscious thought. This suggests that they are unlikely to consider others’ beliefs and strategies in games and are therefore more likely to play Level-1 actions.

Unfortunately, this result is fairly weak. A multivariate logit regression testing for correlations between quiz scores and all five levels simultaneously (Table X) finds marginal significance in the memory test score predicting Level-1 play, but all other coefficient estimates are not significant. The result is similar for the guessing games: The Eye Gaze score is significantly correlated with Level-1 play but the memory score is not, while no other coefficients are significant.<sup>15</sup> Thus, we conclude that there exists some

<sup>15</sup>Ordered logit regressions find no significant estimates on the effect of quiz scores. Again, we expect this occurs because the observed levels are not monotonic in subjects’ underlying strategic sophistication.

	# obs.	Const.	IQ	EyeGaze	Memory	CRT	Takeover
Level-0	4	-13.165 (0.051)	0.306 (0.542)	1.139 (0.110)	0.021 (0.951)	0.080 (0.692)	-0.642 (0.096)
Level-1	40	1.731 (0.319)	0.077 (0.656)	-0.212 (0.118)	<b>-0.176</b> <b>(0.047)</b>	-0.062 (0.316)	0.005 (0.963)
Level-2	31	-2.918 (0.144)	0.097 (0.608)	0.167 (0.292)	-0.004 (0.975)	0.045 (0.476)	-0.008 (0.943)
Level-3	3	-9.705 (0.381)	-0.278 (0.551)	-0.168 (0.647)	1.350 (0.354)	0.192 (0.278)	-0.075 (0.803)
Level-4+	31	-1.179 (0.540)	-0.112 (0.529)	0.214 (0.205)	-0.060 (0.552)	-0.015 (0.820)	-0.021 (0.844)
No Majority Level	65	-	-	-	-	-	-

TABLE X. Multivariate logit estimates of the level played in a strict majority of undercutting games on quiz scores.  $p$ -values of coefficient estimates appear in parentheses.

weak correlation between Eye Gaze and memory scores and the incidence of Level-1 play, but that these five quizzes are generally not useful in predicting a player’s realized level.

Finally, note that players who do not play any one level in a majority of games serve as the omitted category in the multivariate logit regression. We observe 65 such players, more than any in other category. This suggests that the cross-game stability of absolute levels observed within the family of undercutting games is not especially strong; the modal ‘type’ is one who switches his realized level between these games.

### *Responsiveness to Signals About Opponents*

In each game each subject is asked to choose a strategy against a randomly-selected opponent, against the person in the room (other than themselves) with the highest total quiz score, and the person in the room (other than themselves) with the lowest total quiz score. Although quiz scores are not strongly related to levels of play—and the relationship certainly is not linear—they are correlated with total earnings, so we hypothesize that subjects treat quiz scores as proxies for strategic sophistication.<sup>16</sup> In other words, information about quiz scores serves as a signal  $\tau_i$  about the expected level of one’s opponent. How subjects respond to their opponents’ characteristics provides another possible testable prediction for the Level- $k$  model.

<sup>16</sup>Many subjects’ responses to a debriefing questionnaire confirm this hypothesis.

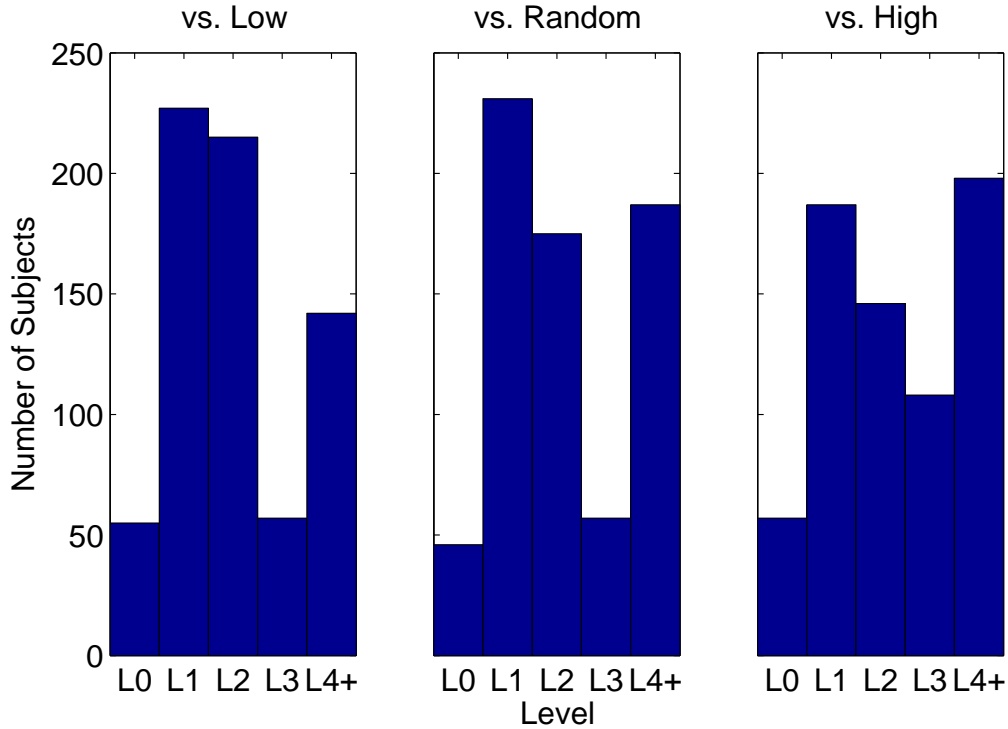


FIGURE VIII. Level distributions by opponent in the four undercutting games.

Figure VIII shows the histogram of choices in the undercutting games for each of the three opponents. Kolmogorov-Smirnov tests for differences between distributions confirm that subjects react differently between the low-score and high-score opponent ( $p$ -value of  $< 0.001$ ) and differently between the random and high-score opponent ( $p$ -value of  $0.007$ ), but not significantly different between the low-score and random opponent ( $p$ -value of  $0.105$ ). Thus, subjects tend to shift levels up against ‘stronger’ opponents, but do not seem to shift down against ‘weaker’ opponents.

Quiz scores are not useful in predicting how much a player shifts their strategies against different opponents. For each subject we measure the average number of levels the subject shifts up (or down) when moving from the low to random, low to high, or random to high opponent. When this average shift is regressed against or correlated with the five quiz scores, no coefficients or correlations are found to be significantly different from zero. Thus, quizzes fail to measure the propensity to adjust play against stronger opponents.

Looking at which subjects do *not* shift strategies yields more informative results. In the undercutting games we identify each subject by their median level. A subject is then

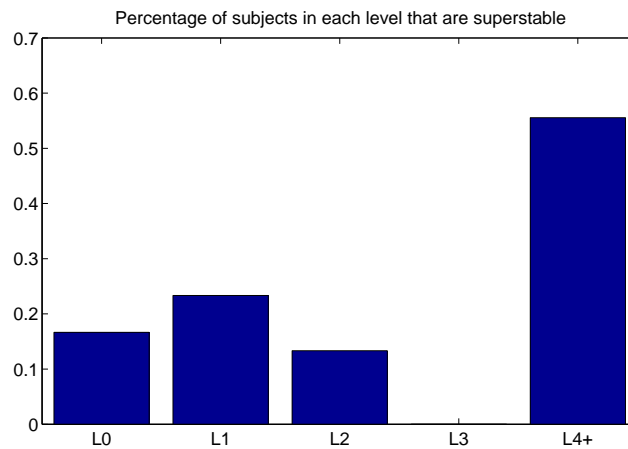


FIGURE IX. For each level  $k$ , the percentage of subjects with a median level of  $k$  that do not choose different strategies against different opponents (undercutting games only).

identified as ‘super-stable’ if they play the same level in all four UGs against all three opponents. The percentage of each subjects in each level that are super-stable is shown in Figure IX. If low-level subjects are constrained by a low capacity then we expect a larger fraction of low-level players to be super-stable. In fact it is the high-level players that are most frequently super-stable. Either these players are playing Nash equilibrium strategies dogmatically in all of these games, or they are high-capacity types whose realized levels are always above four. Given past results on the empirical frequencies of various levels, we suspect the former to be more likely. This suggests a ‘stubborn Nash’ type should be considered as part of a heterogeneous behavioral model.<sup>17</sup>

In summary, we do see some subjects adjusting their realized levels against different opponents, indicating some responsiveness to signals about opponents, but neither the observed levels nor the quiz scores are useful in predicting *which* subjects will make this adjustment.

#### *The Persistence of Players’ Ordering of Games*

An alternative identifying restriction one might impose on the Level- $k$  model is that the ranking of games be consistent between players. Formally, this would require that if

<sup>17</sup>Many authors have included Nash types in heterogeneous behavioral models, but the observation of stubbornness is novel.

	Frequency	i.i.d. Prob.
Standard Guessing Games		
Neither player changes level:	10.8%	8.9%
One player changes level:	42.1%	41.9%
Both change in same direction:	30.3%	24.6%
Players change in opposite directions:	16.8%	24.6%
Opposite dir./same dir. ratio:	0.554	1.00
Undercutting Games		
Neither player changes level:	34.0%	6.6%
One player changes level:	48.7%	38.2%
Both change in same direction:	9.3%	27.6%
Players change in opposite directions:	8.0%	27.6%
Opposite dir./same dir. ratio:	0.860	1.00
Both Families of Games		
Neither player changes level:	9.5%	5.3%
One player changes level:	36.7%	35.6%
Both change in same direction:	28.0%	29.6%
Players change in opposite directions:	25.8%	29.6%
Opposite dir./same dir. ratio:	0.921	1.00

TABLE XI. Observed frequency of game-rank switching among random pairs of subjects between randomly-drawn games, compared to the expected frequency under independently-drawn (i.i.d.) types.

$k_i(\gamma, \tau) \geq k_i(\gamma', \tau)$  for some  $i$  and  $\gamma$  then  $k_j(\gamma, \tau) \geq k_j(\gamma', \tau)$  for all  $j$ . In this way the Level- $k$  model could be thought of a model of (relative) game difficulty or complexity, rather than a model of (relative) strategic sophistication.

Table XI shows the frequency with which a randomly-drawn pair of players changes levels in the same direction when moving between two randomly-chosen games. We first identify cases where one or both players do not change levels from one game to the other; in the remaining two cases the players either change their levels in the same direction or in opposite directions. We compare these frequencies against the expected frequencies if levels were drawn independently from the empirical distribution of types.

In the guessing games we find some support for stability of game orderings. The frequency of opposite-direction switches is 16.8%—less than the 24.6% predicted by randomly-drawn levels—and opposite-direction switches occur roughly half as frequently as same-direction switches. In undercutting games, however, subjects' ordering of games is less correlated. This is mainly due to the stability of absolute levels in these games,

which makes it somewhat rare to find a pair of subjects and games where both subjects strictly switch levels between games. Contingent on that event, however, opposite-direction switches are nearly as frequent as same-direction switches. Using one player's ordering of games is not particularly informative about how another player will play those two games. Clearly, pooling the two families of games gives an intermediate result.

## VIII DISCUSSION

Our goal was to methodically analyze the persistence of strategic sophistication as measured through the Level- $k$  framework that has been used widely to explain behavior in laboratory games. We found only modest support for the notion that behavior conforms to the testable predictions of a model based on strategic sophistication.

Our main finding is that players' Level- $k$  types are often not constant between similar games, and are even less so for games that differ more substantively. This is not simply the result of all subjects' types varying systematically across games, as though some games were uniformly harder than others. Instead, the ordering of subject types often changes across games. Therefore, the notion that the Level- $k$  framework identifies subjects who are persistently "more sophisticated" than others is questionable in our data.

We also find other reasons to question the notion of strategic sophistication. There is little relationship between performance in simple quizzes used in previous studies to measure attributes like strategic intelligence and subjects' type classification in the Level- $k$  framework. While some of these quizzes have correlated with levels of reasoning in prior experiments, we believe that a true measure of strategic sophistication would allow one to predict sophisticated play across a wide variety of games. While there may be some systematic difference between players that conform to a Level-1 type and those that conform to higher levels (based on the ability to think about one's opponent's thoughts), it is harder to find systematic differences between players of higher levels. We also found that what subjects actually play depends on whom they think they are playing against, though we have no systematic prediction of which subjects make this adjustment. Therefore, we find it hard to identify in our experiment who is strategically sophisticated in a manner that would allow us to predict their behavior in novel games.

Moreover, a substantial change to the best-response structure in one class of games yields very little change in behavior. This behavior is inconsistent with the predictions of Level- $k$  models, which depend on the idea that players best respond to *some* beliefs.

Of course, some important caveats are in order. Much work before ours finds support for the Level- $k$  framework as a descriptive theory of behavior in games. Those papers paint a compelling picture of the usefulness of such an approach, by showing that models based on heterogeneous strategic sophistication can account for behavior in a wide variety of contexts. Our experiment employs a different approach, by exploring the extent to which strategic sophistication is consistent across contexts. Our experiment does not discredit the Level- $k$  framework as used by previous researchers—the models are still very useful for describing a particular set of data. Our experiment also uses different subject populations and in some cases different games, meaning that persistent strategic sophistication very well may exist in some of the previous data, even if we find little evidence of it in our experiment.

Our study should perhaps best be viewed as another piece in the complex process of uncovering the behavioral underpinnings of play in laboratory games, rather than a criticism of prior work. Our hope is that these additional insights will help in the formation of a second-generation model of behavior in future research.

#### REFERENCES

- Altmann, S., Falk, A., 2009. The impact of cooperation defaults on voluntary contributions to public goods, university of Bonn Working Paper.
- Ball, S. B., Bazerman, M. H., Carroll, J. S., 1991. An evaluation of learning in the bilateral winner's curse. *Organizational Behavior and Human Decision Processes* 48, 1–22.
- Baron-Cohen, S., 1990. Autism: A specific cognitive disorder of 'mind-blindness'. *International Review of Psychiatry* 2 (1), 81–90.
- Baron-Cohen, S., Jolliffe, T., Mortimore, C., Robertson, M., 1997. Another advanced test of theory of mind: Evidence from very high functioning adults with autism or asperger syndrome. *Journal of Child Psychology and Psychiatry* 38, 813–822.
- Bosch-Domènech, A., Garcia-Montalvo, J., Nagel, R. C., Satorra, A., 2002. One, two, (three), infinity...: Newspaper and lab beauty-contest experiments. *American Economic Review* 92 (5), 1687–1701.
- Bruguier, A. J., Quartz, S. R., Bossaerts, P. L., 2008. Exploring the nature of "trading intuition", California Institute of Technology working paper.
- Camerer, C. F., 2003. *Behavioral Game Theory*. Princeton University Press, Princeton, NJ.

- Camerer, C. F., Ho, T.-H., 1998. Experience-weighted attraction learning in coordination games: Probability rules, heterogeneity and time-variation. *Journal of Mathematical Psychology* 42, 305–326.
- Camerer, C. F., Ho, T.-H., Chong, J.-K., 2004. A cognitive heirarchy model of games. *Quarterly Journal of Economics* 119 (3), 861–898.
- Chen, C.-T., Huang, C.-Y., Wang, J. T.-y., 2009. A window of cognition: Eyetracking the reasoning process in spatial beauty contest games, national Taiwan University working paper.
- Costa-Gomes, M., Crawford, V. P., 2006. Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review* 96 (5), 1737–1768.
- Costa-Gomes, M., Crawford, V. P., Broseta, B., 2001. Cognition and behavior in normal-form games: An experimental study. *Econometrica* 69 (5), 1193–1235.
- Crawford, V. P., 1995. Adaptive dynamics in coordination games. *Econometrica* 63, 103–143.
- Crawford, V. P., Iriberri, N., 2007a. Fatal attraction: Saliency, naivete, and sophistication in experimental “hide-and-see” games. *American Economic Review* 97 (5), 1731–1750.
- Crawford, V. P., Iriberri, N., 2007b. Level- $k$  auctions: Can a nonequilibrium model of strategic thinking explain the winner’s curse and overbidding in private-value auctions? *Econometrica* 75 (6), 1721–1770.
- Devetag, G., Warglien, M., 2003. Games and phone numbers: Do short-term memory bounds affect strategic behavior? 24, 189–202.
- Duffy, J., Nagel, R. C., 1997. On the robustness of behavior in experimental ‘beauty contest’ games. *Economic Journal* 107, 1684–1700.
- Erev, I., Roth, A. E., 1998. Prediction how people play games: Reinforcement learning in games with unique strategy equilibrium. *American Economic Review* 88, 848–881.
- Fehr, E., Schmidt, K. M., 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114 (3), 817–868.
- Frederick, S., 2005. Cognitive reflection and decision making. *Journal of Economic Perspectives* 19, 25–42.
- Georganas, S., 2009. English auctions with resale: An experimental study, university of Bonn working paper.
- Harsanyi, J. C., 1967. Games with incomplete information played by “bayesian” players, i-iii. part i: The basic model. *Management Science* 14, 159–182.

- Ho, T.-H., Camerer, C. F., Weigelt, K., 1998. Iterated dominance and iterated best response in  $p$ -beauty contests. *American Economic Review* 88, 947–969.
- Ivanov, A., Levin, D., Niederle, M., 2008. Can relaxation of beliefs rationalize the winner's curse?: An experimental study, stanford University working paper.
- Keynes, J. M., 1936. *The General Theory of Interest, Employment and Money*. Macmillan, London.
- McKelvey, R. D., Palfrey, T. R., 1995. Quantal response equilibria for normal form games. *Games and Economic Behavior* 10 (1), 6–38.
- Nagel, R. C., 1993. Experimental results on interactive competitive guessing, discussion Paper 8-236, Sonderforschungsbereich 303, Universitat Bonn.
- Rogers, B. W., Palfrey, T. R., Camerer, C. F., 2009. Heterogeneous quantal response equilibrium and cognitive hierarchies. *Journal of Economic Theory* 144, 1440–1467.
- Samuelson, W. F., Bazerman, M. H., 1985. The winner's curse in bilateral negotiations. In: Smith, V. L. (Ed.), *Research in Experimental Economics*. Vol. 3. JAI Press, Greenwich, CT, pp. 105–137.
- Stahl, D. O., Wilson, P. O., 1994. Experimental evidence on players' models of other players. *Journal of Economic Behavior and Organization* 25 (3), 309–327.
- Stahl, D. O., Wilson, P. W., 1995. On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior* 10, 218–254.
- Strzalecki, T., 2009. Depth of reasoning and higher-order beliefs, harvard University Working Paper.
- Wechsler, D., 1958. *The measurement and appraisal of adult intelligence*, 4th Edition. Williams & Wilkins, Oxford, England.