City Research Online

City, University of London Institutional Repository

Citation: Georganas, S., Healy, P.J. and Weber, R. A. (2015). On the persistence of strategic sophistication. Journal of Economic Theory, 159(A), pp. 369-400. doi: 10.1016/j.jet.2015.07.012

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: http://openaccess.city.ac.uk/12715/

Link to published version: http://dx.doi.org/10.1016/j.jet.2015.07.012

Copyright and reuse: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

ON THE PERSISTENCE OF STRATEGIC SOPHISTICATION[†]

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

ABSTRACT. We examine whether the 'Level-k' model of strategic behavior generates reliable cross-game testable predictions at the level of the individual player. Subjects' observed levels are fairly consistent within one family of similar games, but within another family of games there is virtually no cross-game correlation. Moreover, the relative ranking of subjects' levels is not consistent within the second family of games. Direct measures of strategic intelligence are generally not correlated with observed levels of reasoning in either family. Our results suggest that the Level-k model is just one of many heuristics that may be triggered in some strategic settings, but not in others.

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

JEL Classification: C72; C91; D03.

Draft: October 9, 2012

[†]We thank Doug Bernheim, Tilman Börgers, John Kagel, Stephen Leider, John Lightle, Tom Palfrey, Reinhard Selten, Dale Stahl, Joseph Tao-yi Wang, and Matthias Wibral for helpful comments. We are particularly indebted to Colin Camerer and Vince Crawford for their valuable comments and thoughtful suggestions.

[*]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.

I INTRODUCTION

Following a considerable literature demonstrating deviations from Nash equilibrium play (see, for example, Camerer, 2003), behavioral research has sought to model the processes determining individual play and aggregate behavior in experimental games. 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).[1] In this framework—often referred to as *Level-k* or *Cognitive Hierarchy*—players' strategic sophistication is represented by the number of iterations of best response they perform in selecting an action.

In the simplest version of these 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 level is a measure of her strategic sophistication, or, more precisely, her belief about her opponents' strategic sophistication. 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.

The value of the Level-k framework as a *post hoc* descriptive model of the aggregate distribution of actions in laboratory games has been widely documented. There is also evidence that the overall distribution of levels may posses some stability across games (e.g., Camerer et al., 2004), meaning that one might be able to predict the distribution of actions in a novel game based on the distributions in other games. However, 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 a particular individual's estimated level correspond to a persistent trait that can be used to predict play across games? If levels are indicative of strategic sophistication, and if strategic sophistication is an invariant characteristic of a person, then there should exist reliable cross-game patterns in players' observed levels. In this case, estimated levels in one game can be used to predict players' behavior in novel games. Moreover, a player's level may then be predictable using psychometric measures other than actual game play.

[1] 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 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.).

On the other hand, if players' levels appear to be randomly determined from game to game, then one of two negative conclusions must be reached: Either iterative best response is not an accurate description of players' reasoning, or else the model is accurate but players' levels vary from game to game in a manner that is difficult to predict. In either case, knowledge of a player's level in one game provides neither information about their play in another game, nor a useful measure of that person's strategic sophistication in general.[2]

In this paper we test for persistence of individuals' strategic sophistication across games. We begin by identifying several plausible, testable restrictions on cross-game behavior in the Level-k framework. For example, the most stringent testable restriction is that players' levels are constant across all games. A weaker restriction requires only that players' relative levels be invariant, so that a ranking of players based on their levels remains constant across games, even if their absolute levels do not. We then conduct a laboratory experiment in which subjects play several games drawn from two distinct families of games. Within each game, we identify an individual's level in a Level-k framework. We then test whether these observed levels satisfy any of the cross-game restrictions we have identified.

The first family of games consists of four novel matrix games developed for this study, which we refer to as 'undercutting games'. The second family is a set of two-person guessing games studied by Costa-Gomes and Crawford (2006) (henceforth CGC06).[3] Comparing our data in the guessing games to that of CGC06 provides both a consistency check and an avenue to explore the robustness of the Level-k model to variations in experimental protocols, since subjects in the CGC06 experiments were given much more extensive instructions and an understanding test in which they were required to calculate best responses for themselves and for their opponents before starting the experiment.

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 independent notions of strategic sophistication, we elicit several direct measures of strategic intelligence using brief quizzes that are known to identify strategic

[2]We do not suggest that levels must be constant across games for the model to have predictive power. Camerer et al. (2004) and Chong et al. (2005), for example, suggest that levels *will* change in certain situations. Predictive power of the model is maintained if (and only if) those situational changes are predictable.

[3]A two-person guessing game is different from the two-person beauty contest studied by Grosskopf and Nagel (2008); the latter has a (weak) dominant strategy while the former does not.

reasoning ability or general intelligence. We explore the relationship between such measures and subjects' levels 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 opponent who scored highest on the strategic intelligence measures discussed above, and the opponent 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. The key results are summarized as follows:

(1) The *aggregate* distribution of levels is similar to that found in previous studies.
(2) The aggregate distribution of levels is remarkably stable across undercutting games, but quite unstable across two-person guessing games.
(3) Individual levels are moderately persistent within the family of undercutting games, but have no persistence within the family of guessing games.
(4) For any two players, the relative ordering of their levels is stable between undercutting games, but not stable between guessing games.
(5) The quizzes generally fail to predict players' levels in either family of games, though Level-1 play is weakly correlated with a test for autism and poor short-term memory.
(6) Some players adjust strategies against stronger opponents, but neither quiz scores nor levels predict which subjects make this adjustment.
(7) Subjects in our experiment conform exactly to the Level-k predictions much less frequently than in CGC06. This difference is likely due to differences in instructions and CGC06's use of a best-response understanding test. Thus, the model's fit is not robust to experimental protocols.
(8) When a guessing game's payoff function is modified to introduce dominant strategies, aggregate behavior does not significantly change and very few players select a dominant strategy.

Our interpretation of these results is that the congruence between Level-k models and subjects' actual decision processes depends on the context. While in a few instances, such as within the family of undercutting games, we find robust evidence of persistent strategic sophistication, we find much less support when considering the guessing games or other features of the data. One way to interpret our findings is that players have available many heuristics by which they may choose strategies in games, and different environments trigger the use of different heuristics. This is very similar to the model

of behavior put forth by Gigerenzer (2001). Perhaps in our undercutting games—and in many past studies where the Level-k model fit well—the Level-k heuristic was utilized by many of the subjects. In that case, a player's absolute level may appear fairly constant across games. In the guessing games, however, it appears that different heuristics were triggered. Not only were estimated levels highly unstable between guessing games, but changes to the best response function did not yield significant changes in behavior, suggesting that the operating heuristic did not rely critically on best responses.

In the CGC06 guessing game experiments the Level-k model receives stronger support, especially when the choice data is augmented by 'lookup' data indicating which game parameters the subject focused on during the experiment.[4] We believe this difference in model accuracy stems from their lengthier instructions and their best-response understanding test, either of which may trigger the Level-k heuristic in some fraction of the subjects. The Level-k heuristic may not be not triggered as frequently among subjects who receive shorter instructions and no best-response understanding test.

This multiple-heuristics model of strategic thinking implies that researchers should take care in extrapolating the success of any one model to out-of-sample strategic settings. Instead, future work should focus on understanding which heuristics are widely used and which features of a strategic environment trigger the use of different heuristics. We speculate that experimental protocols, training, and experience all have an impact on the choice of heuristic, and that the presentation of a game in matrix form (as in our undercutting games) is more likely to trigger best-response-based heuristics like the Level-k model.[5]

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 choose a guess from a known interval (usually $[0,100]$), and a monetary prize is given to the subject whose guess is closest to

[4]Our data are comparable to the CGC06 'Baseline' and 'Open-Box' treatments. In their 'Trained Subjects' treatments, subject were explicitly paid based on their conformance with equilibrium strategies; we do not consider those data here.

[5]Our undercutting games also have the strategies ordered in a natural way, with higher levels playing lower-numbered strategies. We conjecture that 'shuffling' the matrix would reduce the quality of fit with the Level-k model.

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 and Bosch-Domènech et al., 2002).

Shapiro et al. (2009) test a version of the p-beauty contest where subjects are rewarded for minimizing the distance between their guess and p times the average guess, and also for minimizing the distance between their guess and some unknown state of the world about which they observe a public and a private signal, following Morris and Shin (2002). The Level-k model fits the data well in treatments where the beauty-contest component of subjects' payoffs receives the greater weight, but fits poorly when the state-guessing component of payoffs is emphasized. These results suggest the Level-k model may have a limited domain of applicability.

Burnham et al. (2009) do not explicitly address level-k models, but they provide some experimental results that suggest a relationship between intelligence and levels of reasoning in one version of the p-beauty contest game. Specifically, they show that individuals' choices in a p-beauty contest game are correlated with a direct measure of intelligence (a 20-minute test of cognitive ability). More intelligent players play considerably closer to the Nash equilibrium strategy than players who rate low on the cognitive ability measure, whose average choices are close to 50.

Stahl and Wilson (1994) study Level-k behavior in ten 3×3 matrix games.[6] They find that roughly 25 percent of players are Level-1, 50 percent are Level-2, and 25 percent are Nash-equilibrium 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 four 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 as the level increases. In both studies, many subjects fit strongly into one type, with posterior probabilities of their maximum likelihood type exceeding 0.90. Interestingly, Stahl and Wilson (1995) provide a rare example of a test of individual cross-game stability: They select a subset of nine games, estimate individuals' types from these games, calculate the predicted choice probabilities for the remaining three games for each type, and then estimate the posterior probability that a subject has a particular type. They classify as "stable" those subjects for whom the posterior probability of having the same type is at least 15

[6] 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.

percent. Using this approach, they find that, for 35 of 48 subjects the subject's behavior in the second set of games makes it at least 15 percent likely that the subject has the same type as was estimated in the first set of games. While this represents a valuable instance of researchers exploring the stability of individual types, the approach seems likely to confirm type stability to a greater extent than the more cautious method of estimating types independently in two sets of games, as we do here.

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 τ (after defining the Level-0 distribution) that describes the mean level in the population. They estimate this distribution for a wide range of games. In p-beauty contests, for example, they estimate higher mean levels in more educated populations, in simpler games, and when subjects are asked their beliefs about opponents' play. They also show that the model suffers relatively little loss in likelihood scores when restricting τ to be constant across games, indicating a fair amount of cross-game stability in the aggregate distribution of levels; however, they do not explore individual-level cross game stability.

Chong et al. (2005) describe how one can take an estimated population distribution of levels and assign levels to individuals using a maximum-likelihood technique that requires the distribution of assigned levels to match the original estimated distribution of levels. Their subjects play 22 mixed-equilibrium matrix games in a fixed order. Chong et al. (2005) report a positive correlation between thinking time and levels. Furthermore, average levels are higher in games 12-122 than in games 1–11, indicating a learning-by-doing increase in sophistication over time. In a personal communication, Camerer reported that a regression of individuals' average second-half level on their first-half level yields an R^2 value of 0.37, indicating reasonable predictive power in these games despite the learning-by-doing effect.[7]

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

[7]Our experiment reduces the incidence of learning-by-doing effects by allowing subjects to revise any of their past decisions after making choices in all ten games. In the appendix we study robustness of our results by estimating types for various subsets of games, rather than for each game individually.

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.

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. Again, lookup behavior is used to strengthen type estimation. The estimation is similar to the previous paper, adding uniform play with probability ε 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. Chen et al. (2009) study similar two-person 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 are somewhat more uniform than in past studies. When subjects' data are randomly re-sampled to generate new bootstrapped samples, however, only 8 of 17 subjects receive the same classification in at least 95% of the bootstrapped samples as they did in the original sample. This suggests that roughly half of the subjects are not strongly consistent with any one level across these games.

Crawford and Iriberri (2007a) analyze 'hide-and-seek' 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.

The Level-k model has also been applied successfully to a variety of other games, including incomplete-information betting games (Brocas et al., 2009), betting games and matrix games (Rogers et al., 2009), sender-receiver games augmented with eye-tracking data (Wang et al., 2009) , and cheap-talk games (Kawagoe and Takizawa, 2009). In the field, Level-k has been shown to fit behavior in Swedish lowest-unique-positive-integer lottery games (Ostling et al., 2009) and to explain the fact that movies that were not released to critics before their public opening earn higher revenues (Brown et al., 2010z). A functional MRI study even suggests differences in brain activity between

subjects who exhibit varying degrees of 'strategic sophistication' (Bhatt and Camerer, 2005). For a recent survey, see Crawford et al. (2010).

Finally, many recent papers apply the Level-k concept to study departures from Nash equilibrium play in auctions. For example, Crawford and Iriberri (2007b) apply the Level-k model to auction data from many different experiments and find that in many 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. (2009a) 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.

[ADD GILL & PROWSE, WHO FIND CORRELATION BTWN COGNITIVE ABILITY AND LEVELS IN P-BEAUTY CONTESTS]

III A Formulation of Level-k Models

The usual applications of the Level-k model generally treat it as an *ex post* descriptive model. As such, prior analyses typically omit cross-game or cross-individual testable restrictions, other than perhaps tests of how the aggregate distribution of types varies across games or populations (Camerer et al., 2004, e.g.). In this section we introduce 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 simple 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 her entire hierarchy of beliefs, and that all players believe all others to have strictly lower levels than themselves.[8]

[8]For example, Costa-Gomes et al. (2001), Costa-Gomes and Crawford (2006), Crawford and Iriberri (2007b), and Crawford and Iriberri (2007a) assume that all players with a level of $k > 0$ believe all other

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 $u_i : S \to \mathbb{R}$ and $u_j : S \to \mathbb{R}$. The set of all such two-player games is Γ. When players use mixed strategies $\sigma_i \in \Delta(S_i)$ we abuse notation slightly and let $u_i(\sigma_i, \sigma_j)$ and $u_j(\sigma_i, \sigma_j)$ represent their expected payoffs. In some cases players receive 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 $c_i : \Gamma \to \mathbb{N}_0 := \{0,1,2,\ldots\}$ identifies i's capacity for each game $\gamma \in \Gamma$, and $k_i : \Gamma \times T \to \mathbb{N}_0$ 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, γ, and τ_i.[9] Let Θ be the space of all possible types. Note that c_i does not vary in τ_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 τ_i, however, because the realized 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 $\nu : \mathbb{N}_0 \to \Delta(\mathbb{N}_0)$ such that $\nu(k_i)(\{0,1,\ldots,k_i-1\}) = 1$ for all $k_i \in \mathbb{N}_0$.[10] For example, in Camerer et al. (2004), $\lambda > 0$ is a free parameter and $\nu(k)(l) = (\lambda^l/l!)/\sum_{\kappa=0}^{k-1}(\lambda^\kappa/\kappa!)$ if $l < k$ and $\nu(k)(l) = 0$ otherwise. The function ν is common knowledge and therefore is not included in the description of θ_i. Thus, the k_i component of a player's type completely identifies her beliefs since ν is a function only of k_i; this is a common implicit assumption in the literature.

Behavior in a Level-k model is defined inductively. The Level-0 strategy for each player i in γ is given exogenously as $\sigma_i^0 \in \Delta(S_i)$. If $k_i(\gamma, \tau_i) = 0$ then player i plays σ_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 a best

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 $\{0,1,\ldots,k-1\}$. 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.

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

[10] The simple interpretation of this assumption is that each player believes they are more 'sophisticated' than all of their opponents. An alternative interpretation is that players are aware that they may be less sophisticated than some of their opponents, but they have no model of how more sophisticated players choose strategies. More sophisticated players are then treated as though they are Level-0 players. This second interpretation does suggest that $\nu(k_i)(0)$ should be positive for all k_i, which is inconsistent with the commonly-used assumption that $\nu(k_i)(k_i - 1) = 1$ for all k_i.

response to beliefs $\nu(k)$, given that each level $\kappa < k$ of player j plays σ_j^κ.[11] Formally, σ_i^k for each $k > 0$ is such that for all $s_i' \in S_i$,

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

Finally, we define a 'Nash' type, denoted by $k = N$, whose beliefs are $\nu(N)(N) = 1$. The profile σ_i^N is then the best response to the other player's Nash-type strategy σ_j^N.[12]

When σ_i^k is degenerate (the Level-k strategy is a unique pure strategy) we let s_i^k be the strategy such that $\sigma_i^k(s_i^k) = 1$.

To see how this construction operates, fix a game γ and signal τ_i. If player i's type in this situation is $(c_i, k_i) = (0,0)$ then she plays σ_i^0. If i's capacity is one then her type is either $(1,0)$ or $(1,1)$. In the former case she plays σ_i^0; in the latter case her beliefs are $\nu(1)$, which has $\nu(1)(0) = 1$, and so she plays σ_i^1. If i's type is $(2,2)$ then she has beliefs $\nu(2)$, which puts pre-defined probabilities on her opponent being Level-0 and Level-1. In this case she plays σ_i^2. For any (c_i, k_i) player i's beliefs are $\nu(k_i)$ 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.

Once σ_i^0 and ν 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 $\{\sigma_i^0, \sigma_i^1, \sigma_i^2, \ldots\} \cup \{\sigma_i^N\}$.[13] In many applications, the researcher assumes that each level k plays σ_i^k with noise (usually with a logistic distribution) and then assigns each subject to the level that maximizes the likelihood of their data across all games played.

As specified, a player's level $k_i(\gamma, \tau_i)$ can be any arbitrary function of γ and τ_i. If no structure is imposed on the k_i function then the model is incapable of cross-game or cross-signal predictions; knowing that player i plays Level-2 in one game doesn't provide information about i's level in another game. Our goal is to consider a set of reasonable cross-game or cross-signal testable restrictions on k_i and explore which (if any) receive empirical support. Understanding which restrictions on k_i apply will then lead to an understanding of the out-of-sample predictions that can be made through this model. If no restrictions on k_i can be found then no out-of-sample predictions can be made for an individual.

[11] If there are multiple pure-strategy best responses then σ_i^κ can be any distribution over those best responses, and that distribution is assumed to be known by all higher levels.

[12] As is standard, we assume $\nu(k)(N) = 0$ for all $k \neq N$. If multiple Nash equilibria exist then multiple Nash types could be defined, but all of our games have a unique Nash equilibrium.

[13] If σ^0 is not restricted then there are no testable predictions; letting σ^0 equal the empirical distribution of strategies provides a perfect fit.

Examples of possible restrictions on k_i that we can test using our experiments are:

(1) **Constant:** $k_i(\gamma, \tau_i) = k_i(\gamma', \tau_i')$ for all i, γ, γ', τ_i, and τ_i'.
(2) **Constant Across Games:** $k_i(\gamma, \tau_i) = k_i(\gamma', \tau_i)$ for all i, γ, γ', and τ_i.
(3) **Constant Ordering:** If $k_i(\gamma, \tau) \geq k_j(\gamma, \tau)$ for some γ and τ then $k_i(\gamma', \tau') \geq k_j(\gamma', \tau')$ for all γ' and τ'.
(4) **Responsiveness to Signals:** For every γ and i there is some τ and τ' such that $k_i(\gamma, \tau) > k_i(\gamma, \tau')$.
(5) **Consistent Ordering of Games:** For any τ, if $k_i(\gamma, \tau) \geq k_i(\gamma', \tau)$ for some i, γ and γ', 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.[14]

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, for at least some players, is assumed.

The last restriction listed implies that the observed levels can be used to order the games in Γ. If, at some fixed signal, all players play a lower level in γ' than in γ then it can be inferred that γ' is a more difficult or complex game. This enables future out-of-sample predictions, since a player who subsequently plays Level-2 in γ can be expected to play a lower level in γ'.

It is certainly easy to imagine plausible functions 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

[14]Cross-game learning may still generate violates 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 1	10 -10	0 0	0 0	0 0	0 0	-11 0
2	-10 10	0 0	10 -10	0 0	0 0	0 0	0 0
3	0 0	-10 10	0 0	10 -10	0 0	0 0	0 0
4	0 0	0 0	-10 10	0 0	10 -10	10 -10	10 -10
5	0 0	0 0	0 0	-10 10	0 0	0 0	0 0
6	0 0	0 0	0 0	-10 10	0 0	0 0	0 0
7	0 -11	0 0	0 0	-10 10	0 0	0 0	-11 -11

FIGURE I. Undercutting game 1 (UG1).

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

FIGURE II. Undercutting game 2 (UG2).

or cross-signal restrictions, then the model cannot be used to make out-of-sample predictions about individual behavior. Thus, the predictive power of the model hinges on the presence of at last 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	3	4	5	6	7	8	9
1	1 1	10 -10	0 0	0 0	0 0	0 0	0 0	0 0	-11 0
2	-10 10	0 0	10 -10	0 0	0 0	0 0	0 0	0 0	0 0
3	0 0	-10 10	0 0	10 -10	0 0	0 0	0 0	0 0	0 0
4	0 0	0 0	-10 10	0 0	10 -10	0 0	0 0	0 0	0 0
5	0 0	0 0	0 0	-10 10	0 0	10 -10	0 0	0 0	0 0
6	0 0	0 0	0 0	0 0	-10 10	0 0	10 -10	10 -10	10 -10
7	0 0	0 0	0 0	0 0	0 0	-10 10	0 0	0 0	0 0
8	0 0	0 0	0 0	0 0	0 0	-10 10	0 0	0 0	0 0
9	0 -11	0 0	0 0	0 0	0 0	-10 10	0 0	0 0	-11 -11

FIGURE III. Undercutting game 3 (UG3).

	1	2	3	4	5	6	7
1	1 1	10 -10	0 0	0 0	0 0	0 0	-11 0
2	-10 10	0 0	10 -10	0 0	0 0	0 0	0 0
3	0 0	-10 10	0 0	10 -10	0 0	0 0	0 0
4	0 0	0 0	-10 10	0 0	30 -30	10 -10	10 -10
5	0 0	0 0	0 0	-30 30	0 0	0 0	0 0
6	0 0	0 0	0 0	-10 10	0 0	0 0	0 0
7	0 -11	0 0	0 0	-10 10	0 0	0 0	-11 -11

FIGURE IV. Undercutting game 4 (UG4).

$\{1,2,\ldots,m,\ldots,n\}$. 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 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. The cases where a player loses 11 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 $m = 4$ ($s^1 = 4$) as it maximizes the sum of row payoffs, Level-2 types play 3 ($s^2 = 3$), Level-3 types play 2 ($s^3 = 2$), and all higher levels play the equilibrium strategy of 1

($s^N = 1$). This enables a unique identification of a player's level (up to Level-4) 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 $e_i = |s_i - p_i s_j|$.

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 $s_i^1 = p_i(a_j + b_j)/2$; if this is not attainable then the Level-1 player will select the nearest endpoint of her interval. A Level-2 type will play $s_i^2 = p_i s_j^1$ (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).

☐ jealous ☐ panicked ☐ arrogant ☐ hateful

☐ aghast ☐ fantasizing ☐ impatient ☐ alarmed

FIGURE V. Sample questions from the Eye Gaze test.

V EXPERIMENTAL DESIGN

In total, 116 undergraduate students from Ohio State University participated as subjects in these experiments. After reading through the experiment instructions, 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.[15] Similar tests of cognitive ability have been shown to correlate with higher levels of reasoning on p-beauty contests (Burnham et al., 2009).

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 autism or Asperger's

[15]See http://www.mensa.org/workout2.php

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 τ: 0.248) and significant (p-value: 0.010). Camerer et al. (2004) use this observation as a plausible justification for their assumption that the the relative frequencies of two consecutive levels k and $k-1$ ($f(k)/f(k-1)$) is declining in k, which then motivates their restriction to Poisson distributions of levels.

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 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 are 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.[16] The sum of the four quiz scores was calculated

[16]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.

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.

for each player. Players are given no feedback about any player's absolute or relative performance on the quizzes until the end of the experiment, at which point they learn only their own total quiz score.

After completing the quizzes, the subjects play ten games against varying opponents. The first four games are undercutting games and the last six are guessing games. The parameters of each game are given in Table I and Figures I–IV. The final three guessing games are identical to the first three, with the players' roles reversed. As in Costa-Gomes and Crawford (2006), this allows players to play both roles and also allows subjects' decisions in GG5, for example, to be matched with another subject's player-1 decision in GG8 to determine payoffs.

In each game subjects are 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 are made without feedback. After making these three choices in all ten games, players learn that they could 'loop back' through the games and revise their choices if desired. This can be done up to four times, for a total of five iterations through the ten games, all without feedback. Once subjects finish all five iterations—or decline the opportunity to loop back—their play is recorded, four of their choices are randomly selected (two from the undercutting games and two from the guessing games), and they are matched with another player and paid for their decisions. Subjects earning less than \$6 (the standard show-up fee) are paid \$6 for their time. Subjects earned an average of \$24.85 overall.

VI Data Analysis Procedure

Each subject plays ten games, each against three different opponents, for a total of thirty game-play observations per subject. We employ three signals ($T = \{\tau^{\text{LO}}, \tau^0, \tau^{\text{HI}}\}$) indicating whether the current opponent has the lowest quiz score, is randomly selected, or has the highest quiz score. As in CGC06 (and others), we focus on the case where $\nu(k)(k-1) = 1$ for all $k > 0$ and σ_i^0 is uniform over S_i. The games are chosen so that the estimated levels (or, at least, players' relative rankings of levels) are fairly robust to these assumptions. Furthermore, the guessing game parameters are chosen from among the CGC06 parameters to maximize the distance between any two levels' predicted strategy choices; this helps to minimize the error in subjects' level estimates.

To each subject i, signal τ, and set of games $G \subseteq \Gamma$ we can estimate a level $k_i(G,\tau)$ using a simple maximum-likelihood approach that follows closely CGC06. Specifically, for each level k we define a likelihood function $L(s_{i\gamma\tau}|k,\lambda,\varepsilon)$ that follows a logistic response with sensitivity $\lambda \geq 0$ based on $\nu(k)$ with a 'spike' of size $\varepsilon \in [0,1]$ at $s_{i\gamma}^k$, if $s_{i\gamma}^k$ is well-defined.

Formally, let $\bar{s}_{i\gamma\tau}$ be the value of $s_{i\gamma\tau}$ rounded to the nearest integer, and let $I_i(s_{i\gamma\tau},k)$ be an indicator function that equals one if $\bar{s}_{i\gamma\tau} = s_{i\gamma}^k$, where s_i^k denotes the Level-k strategy for player i in game γ.[17] $I_i(s_{i\gamma\tau},k)$ equals zero otherwise. Thus, $I_i(s_{i\gamma,\tau},k) = 1$ indicates that i played exactly the Level-k strategy, allowing for rounding. The likelihood function for $k \neq 0$ is then given by

$$L(s_{i\gamma\tau}|k,\lambda,\varepsilon) = \varepsilon I_i(s_{i\gamma\tau},k) + (1-\varepsilon)(1-I_i(s_{i\gamma\tau},k))\left(\frac{\exp\left(\lambda\sum_\kappa u_i(s_{i\gamma\tau},\sigma_j^\kappa)\nu(k)(\kappa)\right)}{\int_{S_i}\exp\left(\lambda\sum_\kappa u_i(z_i,\sigma_j^\kappa)\nu(k)(\kappa)\right)dz_i}\right).$$

For $k = 0$ we set $L(s_{i\gamma\tau}|0,\lambda,\varepsilon)$ equal to $\sigma_{i\gamma}^0$, which is assumed to be the uniform distribution over S_i.

For any set of games $G \subseteq \Gamma$, denote i's strategies given signal τ by $s_{iG\tau} = (s_{i\gamma,\tau})_{\gamma\in G}$. For each level $k \in \mathbb{N}_0 \cup \{N\}$, the maximum likelihood of observing $s_{iG\tau}$ is given by

$$L^*(s_{iG\tau}|k) = \max_{\lambda>0,\varepsilon\in[0,1]} \prod_{\gamma\in G} L(s_{i\gamma\tau}|k,\lambda,\varepsilon).$$

In practice, we search over a non-uniform grid of 122 possible values for λ and a uniform grid of 19 possible values for ε. The maximum-likelihood level for these observations is then given by

$$k_i(G,\tau) = \arg\max_{k\in\mathbb{N}_0\cup\{N\}} L^*(s_{iG\tau}|k).$$

[17] All strategies are integers in the undercutting games, in which case $\bar{s}_{i\gamma\tau} = s_{i\gamma\tau}$.

Our games and our model of noisy play are such that the maximum-likelihood level is generically unique. Given that levels greater than three are very rarely observed in past data, we only calculate likelihood values for $k \in \{0,1,2,3,N\}$.

We consider two types of analyses. First, we estimate for each subject one level for all undercutting games ($G = \{1,\ldots,4\}$) and another level for all undercutting games ($G = \{5,\ldots,10\}$). This enables us to compare stability of levels across families of games. This pooling of several games per estimate also matches the standard procedure for estimating levels in the literature. Second, we estimate for each subject a level in *every* game ($G = \{\gamma\}$). This enables us to compare stability of levels within each family of games. [18] In the appendix we also explore intermediate cases where two or three games per estimate are used.

As a robustness check, we apply our procedure to CGC06's data, pooling all games to generate a single estimated level per subject (as in their paper), and find exact subject-by-subject agreement between our estimated levels and theirs. For the case of $|G| = 1$, we also estimate levels in the guessing games by eliminating ε and $I_i(s_{i\gamma\tau},k)$, setting $\lambda = 1.33$ (the average estimated value of λ in CGC06 using only subjects' guesses), and then assigning a k to each observation as described above. Under this new procedure, 85.5% of observations receive the same level assignment as in our original procedure. Roughly half of the observations whose level changes became Level-0 observations, implying their likelihood value simply falls below the uniform distribution likelihood. None of the key results of the paper change under these alternative estimates.

VII Results

Result 1: Aggregate Distributions of Levels

The distribution of levels for each game is shown in Table II, along with the distributions of levels when all undercutting games are pooled, and when all guessing games are pooled. We also take the single-game estimates and provide in Figure VI the average distribution for all ten games, the four undercutting games, and the six guessing games. These distributions represent fairly typical distributions of estimated levels: Level-0 is observed fairly infrequently, Level-1 is the modal type, and higher levels are observed less frequently. The distribution for guessing games is similar to the distribution found

[18] In this case $k_i(G,\tau)$ represents an assignment rule rather than an econometric estimate since only one observation is used for each 'estimate' and no standard errors can be calculated.

Game	L0	L1	L2	L3	Nash
UG1	7.76%	32.76%	19.83%	10.34%	29.31%
UG2	7.76%	32.76%	22.41%	7.76%	29.31%
UG3	5.17%	27.59%	18.10%	5.17%	43.97%
UG4	6.03%	31.03%	29.31%	5.17%	28.45%
UGs Pooled	4.31%	28.45%	26.72%	5.17%	35.34%
GG5	6.03%	70.69%	9.48%	12.07%	1.72%
GG6	0.86%	65.52%	17.24%	11.21%	5.17%
GG7	43.10%	37.07%	13.79%	1.72%	4.31%
GG8	6.90%	39.66%	24.14%	21.55%	7.76%
GG9	5.17%	42.24%	23.28%	4.31%	25.00%
GG10	9.48%	38.79%	24.14%	19.83%	7.76%
GGs Pooled	1.72%	50.00%	10.34%	10.34%	27.59%

TABLE II. Frequency of levels in each game, and when pooling each family of games.

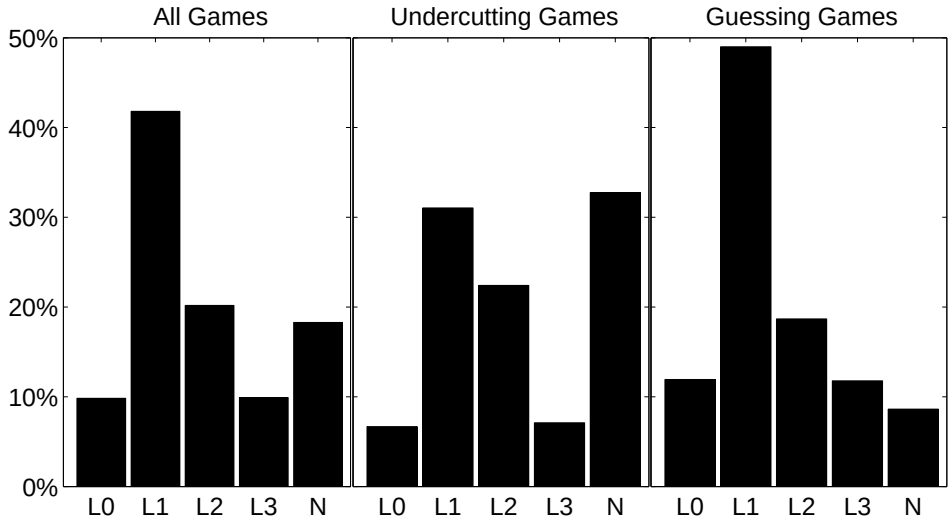

FIGURE VI. Distributions of (single-game) levels, aggregated across all ten games, the four undercutting games, and the six guessing games.

by CGC06 (see Section VIII). We do find that Nash (and higher-than-Level-3) play in our undercutting games is noticeably higher than what is found in most other games.

	L1	L2	L3	Nash	Sum
GG5	8.6%	0%	0%	0.9%	9.5%
GG6	7.8%	1.7%	0%	2.6%	12.1%
GG7	14.7%	0.9%	0%	1.7%	17.2%
GG8	1.7%	0%	0%	6.0%	7.8%
GG9	6.3%	2.6%	0%	20.7%	29.3%
GG10	4.3%	5.2%	0%	0%	9.5%
Average	7.2%	1.7%	0%	5.3%	14.22%

TABLE III. Frequency of exact conformity with the Level-k predictions in the guessing games.

Within the family of undercutting games, the distribution of types is generally stable across games. In all four games, there is a high proportion of L1, L2 and Nash behavior, and relatively little behavior corresponding to L0 and L3.[19]

Within the guessing games, however, distributions of levels vary substantially from one game to the next. For example, the fraction of Level-0 play jumps from 0.68% in guessing game 6 (GG6) to 43.10% in GG7. The fraction of Level-1 play nearly doubles from GG7 to GG5. Nash play ranges from 1.72% in GG5 to 25% in GG9. This suggests that either the Level-k model lacks descriptive power in these games, or else players' levels shift substantially between games.

Result 2: Exact Hits in Guessing Games

Table III reports the percentage of guessing game observations that exactly conform to one of the four (non-zero) levels' predictions, after rounding. In total, 14.22% of observations conform exactly to one of these four levels. This is clearly greater than the 0.7% frequency which would occur if actions were random with a uniform distribution. The most frequently-observed exact hit is the Level-1 action, in which players best respond to the midpoint of their opponent's interval. Nash actions are observed slightly less frequently, though the bulk of these observations come from GG9, where the Nash action is the lower endpoint of the players' strategy space.

An exact-hit rate of 14.22% indicates that, if the Level-k model is descriptive in these games, a relatively high frequency of of 'trembling' is necessary to organize our data. In the 86% of observations without exact hits, the logistic likelihood function must be used to identify a player's type. When pooling all six guessing games, only 14.7% of subjects

[19]Level-0 is necessarily under-counted here, since a proportion of all observed actions should be coming from Level-0 players. Although this cannot be corrected at an individual level, the aggregate frequency can be adjusted. The result simply shifts mass uniformly from the higher levels down to L0.

had multiple exact hits on any given level; the level of the remaining 85.3% must be estimated via the logistic likelihood function. Consequently, the Level-k predictions depend on the exact econometric specification of trembles.

By contrast, 48.9% of the observations in CGC06's data exactly correspond to one of the four levels' predictions. Given similarities in experimental designs and subject pools, we conjecture that this difference is due to differences in experimental instructions and their use of a best-response understanding test; we discuss these procedural differences in Section VIII.

Visualizing Model Fit in Guessing Games

The fit of the Level-k model in a given guessing game can be visualized by plotting a histogram of actions along with likelihood functions for each of the five possible levels. This is done for each game in Figure VII. For simplicity, the likelihood functions are all plotted assuming $\lambda = 1$ and $\varepsilon = 0$. The label for each level appears below its likelihood function's peak, and the Level-0 likelihood appears simply as a uniform distribution over the strategy space. The range of dominated strategies for each game (if any) appears as a dashed line labeled DOM. For any action on the horizontal axis, the assigned level is that whose likelihood function is greatest at that point, given that λ is chosen optimally for each level.

Before analyzing fit, we note two mathematical regularities that arise with the logistic specification. First, the Level-1 likelihood function is much flatter than that of the higher levels. This is because its beliefs are uniform, making deviations from perfect best response less costly in terms of expected loss to the player. Higher levels, by contrast, have degenerate beliefs. Deviations from best response are significantly more costly. If one estimates the Level-k model with randomly-generated data, the Level-1 type will typically be the modal type because of this discrepancy. In other words, the fact that many authors identify the Level-1 type is the most frequently-observed could be an artifact of the logistic specification.

Second, levels whose actions are at the boundary of the strategy space receive nearly double the likelihood for nearby strategies than do levels with interior actions. This is because the trembles beyond the boundary are truncated, and the truncated probability mass is distributed among strategies within the boundaries. For example, in GG8, players who choose actions closer to the Level-3 prediction may still be categorized as Nash types because the Nash likelihood function is amplified by truncation much more than the Level-3 likelihood function. Similar phenomena occur in GG7 and GG9. This

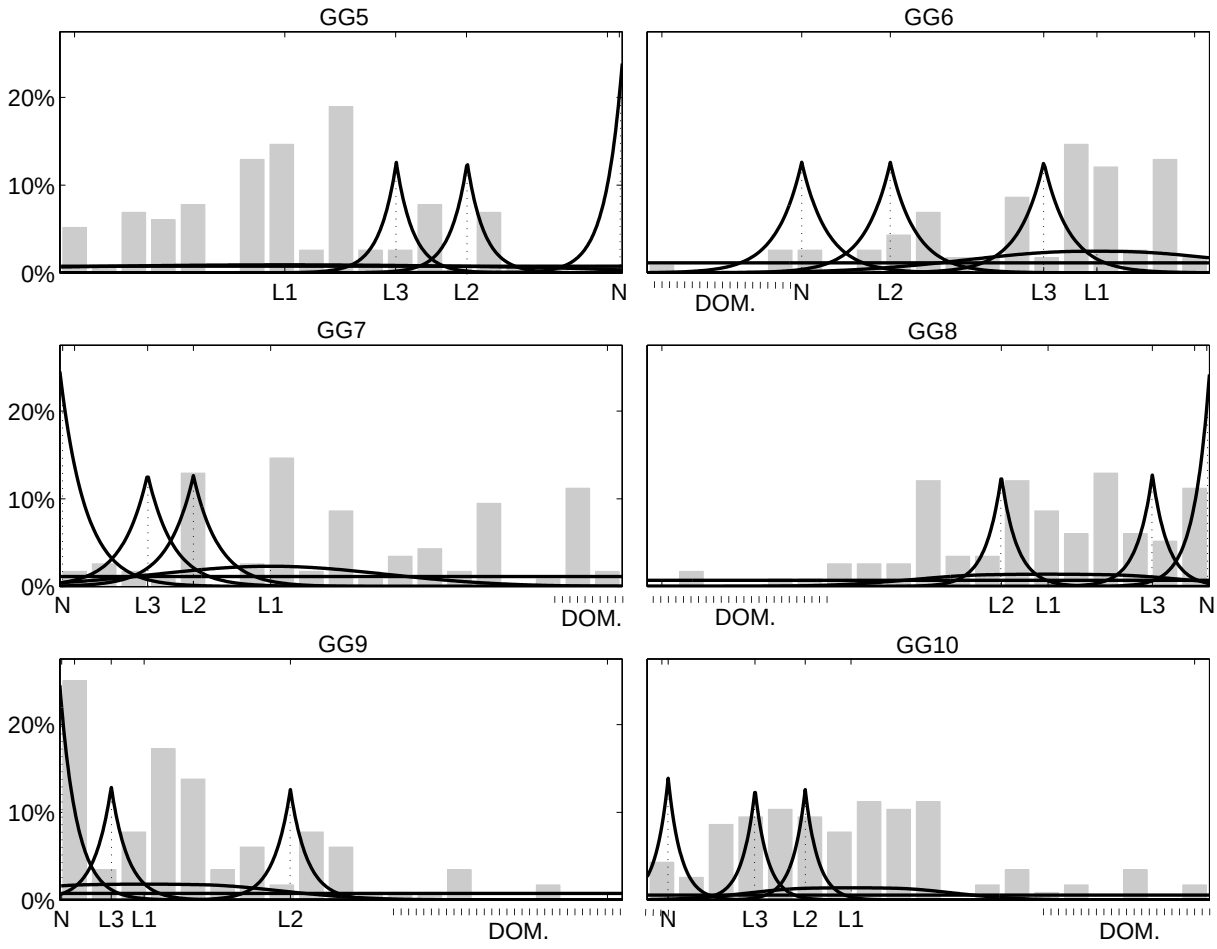

FIGURE VII. Histograms of actions in each guessing game, with likelihood functions for each level (assuming $\lambda = 1$).

is visible in Figure VII. Since Nash types are the only types whose predictions lie at the boundaries, random data will generate relatively larger frequencies of Nash types than Level-3 types. Again, this is consistent with our results.

If the Level-k model fits well, peaks in the histograms should align with peaks in the likelihood functions. Quality of fit clearly differs by game, as was shown in the game-by-game estimated level distributions in Table II. The high proportion of Level-0 types in GG7 is due to players whose action lies in the upper half of the strategy space, while all levels' predictions lie in the lower half. The large frequency of Level-1 types in GG5 comes from that type having a flat likelihood function that captures several peaks in the data. The jump in Nash types in GG9 is due to a large number of players choosing the lower endpoint of the strategy space. If these players are truly using equilibrium logic,

From ↓ To →	L0	L1	L2	L3	Nash
L0	**43.0%**	22.6%	7.5%	9.7%	17.2%
L1	4.9%	**59.7%**	14.6%	4.4%	16.4%
L2	2.2%	20.2%	**57.1%**	9.0%	11.5%
L3	9.1%	19.2%	**28.3%**	18.2%	25.3%
Nash	3.5%	15.6%	7.9%	5.5%	**67.5%**
Overall	6.7%	31.0%	22.4%	7.1%	32.8%

TABLE IV. Markov transition between single-game levels within the four undercutting games.

From ↓ To →	L0	L1	L2	L3	Nash
L0	8.7%	**48.2%**	18.1%	12.3%	12.8%
L1	11.7%	**53.1%**	16.8%	11.2%	7.1%
L2	11.5%	**44.2%**	27.4%	10.0%	6.9%
L3	12.4%	**46.6%**	15.9%	13.2%	12.0%
Nash	17.7%	**40.3%**	15.0%	16.3%	10.7%
Overall	11.9%	**49.0%**	18.7%	11.8%	8.6%

TABLE V. Markov transition between single-game levels within the six guessing games.

From ↓ To →	L0	L1	L2	L3	Nash
L0	0.0%	**60.0%**	0.0%	20.0%	20.0%
L1	6.1%	**42.4%**	6.1%	9.1%	36.4%
L2	0.0%	**51.6%**	16.1%	9.7%	22.6%
L3	0.0%	*33.3%*	*33.3%*	0.0%	*33.3%*
Nash	0.0%	**56.1%**	7.3%	12.2%	24.4%
Overall	1.7%	50.0%	10.3%	10.3%	27.6%

TABLE VI. Markov transitions from the pooled undercutting games to the pooled guessing games.

then most are only doing so in this one game; the frequency of Nash play is much lower in the other five games.

Result 3: Persistence of Absolute Levels

Subjects' levels are fairly stable within the family of undercutting games: Thirty-nine percent of the subjects play the same level in all four undercutting games, and 63 percent play the same level in at least three of the four games. The distribution of levels for these 'stable' players is similar to the aggregate distribution of levels from Figure VI. In the guessing games, however, only six percent play the same level in all six games, and only

37 percent (43 subjects) play the same level in at least four of the six games. Of those 43 subjects, 37 are assigned the Level-1 type, five are assigned the Level-2 type, and one is assigned the Level-3 type in at least four games.

To examine the hypothesis that levels are constant across games ($k_i(\gamma, \tau^0) = k_i(\gamma', \tau^0)$ for all γ and γ'), we generate a Markov transition matrix by selecting all pairs of games (played against random opponents) and, for each level of one game, calculating the frequency with which a subject moves to each level in the other game. Tables IV and V show these transition matrices for the single-game levels in 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 V shows that this result is consistent with levels in each game being independent random draws from the overall distribution of estimated levels (see Figure VI).

As a measure of the stability of levels across games, consider the prediction accuracy of the Level-k model assuming k_i is constant. This is simply the probability that a player plays the same level in two different games. We refer to this probability as the *constant-level prediction accuracy*, or *CLPA*. Mathematically, the CLPA equals the main diagonal of the Markov matrix weighted by the overall probability of each level. If types are constant then the main diagonal entries are all one, as is the CLPA. If types are randomly drawn then each row of the Markov matrix equals the overall distribution, and so the CLPA is simply the sum of squared overall probabilities.

In the undercutting games, the overall frequencies of the levels (given in the last row of Table IV) would imply a CLPA of 26.3% if types were randomly drawn. In fact we observe a CLPA of 57.6%, indicating substantially stronger predictive power than if types were purely random, though still far from the 100% CLPA if levels were truly constant.

To test whether these data yield prediction accuracy greater than that of randomly-drawn levels, we generate 10,000 samples of randomly-drawn levels, with each sample drawn independently using the overall distribution from Table IV. For each sample we calculate the CLPA, generating an approximate distribution of CLPA values under the null hypothesis of random levels.[20] The *maximum* CLPA among the random samples is 33.7%—far less than the observed 57.6%—so we conclude that the null hypothesis is rejected in the undercutting games with a p-value of less than 0.001.

[20]The Monte Carlo simulations use the same number of players as in our experiments. In each game, the distribution of randomly-drawn levels is taken to be the overall distribution of levels for the family of games. Each draw of each level in each game is independent.

Levels are far less persistent in the guessing games (Table V). The realized prediction accuracy (CLPA) is 34.7%. The expected CLPA under randomly-drawn levels is 31.1%. A 10,000-sample Monte Carlo simulation of randomly-generated levels rejects the null hypothesis that the realized CLPA was generated by randomly-drawn levels, with a p-value of 0.003. The absolute magnitude of the difference (34.7% versus 31.1%), however, implies little real gain in predictive accuracy over the assumption of random levels.

When comparing the pooled undercutting game levels to the pooled guessing game levels (Table VI), the CLPA drops to 27.3%, which is actually less than the 29.4% CLPA under the null hypothesis of independent, randomly-drawn levels. In other words, a slight negative correlation is observed in these data. The null is clearly not rejected (p-value 0.68).

The cross-game (or cross-family) correlations can be also be tested statistically for any pair of games by calculating the Cramér correlation coefficient for categorical data (see Siegel and Castellan, 1988, p.225) and comparing it against the null hypothesis of independently-drawn levels, which would give an expected Cramér correlation of zero. For the four undercutting games there are 6 possible unique game pairs in which to test cross-game correlations, each giving a different p-value. The smallest of these correlations is 0.366, with a p-value less than 0.001. The hypothesis of independently-drawn levels is clearly rejected, regardless of which pair of games is used. For the six guessing games there are 15 possible game pairs to test. Of those, 11 give p-values above 0.05, where the null hypothesis of independently-drawn types would not be rejected.[21] Evidence for cross-game correlations appears weak, and limited only to certain pairs of games.

When comparing between the two families of games using pooled-game estimates (Table VI), the null hypothesis of independently-drawn types again cannot be rejected, with a Cramér correlation of only 0.177 and a p-value of 0.562. Thus, it is reasonable to model a player's level in the pooled guessing games as being independent of their type in the pooled undercutting games.

We conclude that estimated levels can reasonably be modeled as constant within certain families of similar games, but not within other families. This suggests that Level-k thinking is applied robustly in some settings, but not in others. Little guidance is currently available as to which families of games will trigger the Level-k heuristic and which will not. In short, using a player's level in one game to predict their action in

[21]The four guessing-game pairs where independence is rejected are $\{5,7\}$, $\{5,8\}$, $\{6,7\}$, and $\{7,10\}$. The p-values in these tests are clearly not independent, and therefore cannot be aggregated to form a single hypothesis test.

	Data	iid Levels
Undercutting Games		
Switch Frequency:	13.2%	27.1%
Non-Switch Frequency:	45.3%	27.1%
Switch Ratio:	0.29	1.00
Guessing Games		
Switch Frequency:	19.9%	23.8%
Non-Switch Frequency:	22.2%	23.8%
Switch Ratio:	0.89	1.00
Pooled UGs vs. Pooled GGs		
Switch Frequency:	25.0%	24.9%
Non-Switch Frequency:	22.7%	24.9%
Switch Ratio:	1.10	1.00

TABLE VII. Observed frequency with which two players' levels strictly switch their ordering, compared to the expected frequency under independent, randomly-drawn levels.

another may be a futile exercise without further information about how the Level-k heuristic is triggered.

Result 4: Persistence of Relative Levels

To examine the frequency with which the ordinal ranking of players' levels changes between games, we consider each possible pair of two games and each possible pair of two players and measure the frequency with which the strictly higher-level player in one game becomes the strictly lower-level player in another ($k_i(\gamma,\tau) > k_j(\gamma,\tau)$ but $k_i(\gamma',\tau) < k_j(\gamma',\tau)$). We refer to this as the 'switch frequency'. This is compared against the 'non-switch frequency', which captures the frequency with which the same player has a strictly higher level in both games ($k_i(\gamma,\tau) > k_j(\gamma,\tau)$ and $k_i(\gamma',\tau) > k_j(\gamma',\tau)$). These will typically sum to less than one, since observations where the two players have the same level in either game are excluded. The 'switch ratio' is the switch frequency divided by the non-switch frequency; this has an expected value of one under the null hypothesis of independently-drawn levels. Under the Level-k model with stable relative levels, the ratio will equal zero.[22]

[22]In practice, the switch ratio would not exactly equal zero since some Level-0 players would be incorrectly identified as higher-level players. This misallocation of levels would occur randomly, generating some apparent switching in the levels when none truly exists. Our simulations suggest the actual switch ratio would be around 0.09 using our overall level distributions; see the appendix for details.

The switch frequency, non-switch frequency, and switch ratio are reported in Table VII for the four undercutting games, the six guessing games, and when comparing the pooled undercutting games to the pooled guessing games. The last column shows the predicted values under the null hypothesis of independently-drawn levels.

Since absolute levels in the undercutting games are fairly stable, we expect similar persistence in subjects' relative levels. This is the case: Non-switching pairs are observed more than three times more frequently than switching pairs, giving a switch ratio of 0.29. In one thousand Monte Carlo simulations of switch ratios for randomly-generated levels (using the undercutting game single-game distributions), the smallest observed switch ratio is 0.77, indicating a clear rejection of the null hypothesis at a 0.001 level.

In the guessing games, however, switching occurs nearly as frequently as non-switching, with a switch ratio of 0.89. The Monte Carlo simulation of one thousand switch ratios using randomly-drawn levels from the single-game distributions yields a marginal p-value of exactly 0.05. Thus, we do not conclusively reject the null hypothesis of randomly-drawn levels in the guessing games when comparing switch ratios.

When the families of games are pooled to generate a single level estimate per family, switching actually occurs more frequently than non-switching. In other words, if Anne exhibits a higher level than Bob in the (pooled) undercutting games, then in the guessing games it is more likely that Bob will exhibit the higher level. In our Monte Carlo simulation with randomly-drawn levels, only 2.3% of simulated switch ratios exceeded 1.10. We can therefore reject the null hypothesis of independently-drawn levels in favor of *negatively* correlated relative levels. This is consistent with our earlier observation that absolute levels are negatively correlated across families of games.

Overall, we conclude that no extra predictive power is gained by considering relative levels instead of absolute levels. Assuming k_i is constant for each subject performs roughly as well as assuming the ordering of k_i across subjects is constant.

Result 5: 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. Observed correlations between scores on the various quizzes are surprisingly weak. IQ, memory, and CRT scores all appear to be positively correlated, though their estimated Spearman rank correlation coefficients all have p-values between 0.05 and 0.10. No other correlations are statistically significant.

	Const.	IQ	EyeGaze	Memory	CRT	Takeover
Coefficient	39.179	0.620	0.250	-0.275	**0.679**	-0.229
p-value	(<0.001)	(0.204)	(0.318)	(0.215)	**(0.002)**	(0.261)

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

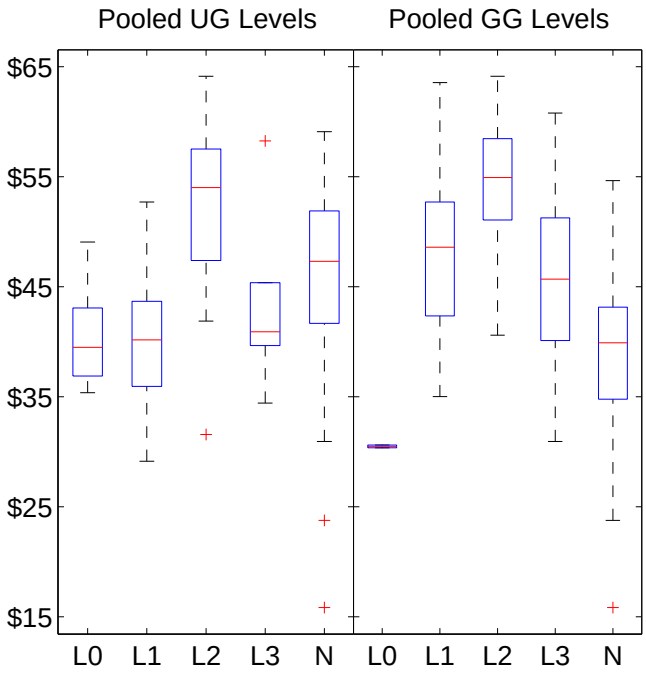

FIGURE VIII. Box plots of total expected earnings distributions for each level, using estimated levels from the pooled undercutting games and estimated levels from the pooled guessing games.

For each subject we calculate their expected earnings in each game when playing their chosen action against the frequency distribution of actions of all possible opponents in that game.[23] The correlation between subjects' total expected earnings across all ten games and the sum of their five quiz scores is positive but not statistically significant (Spearman correlation of 0.172 with p-value 0.064). Regressing total expected earnings on each quiz (Table VIII) reveals that only the Cognitive Reflection Test (CRT) score is correlated with expected earnings.

Although higher levels may appear to be more sophisticated, they do not earn more money. Indeed, Level-2 is the most profitable type since most subjects are estimated

[23]Here, 'all possible opponents' refers to all 115 other subjects, not just those in the same session.

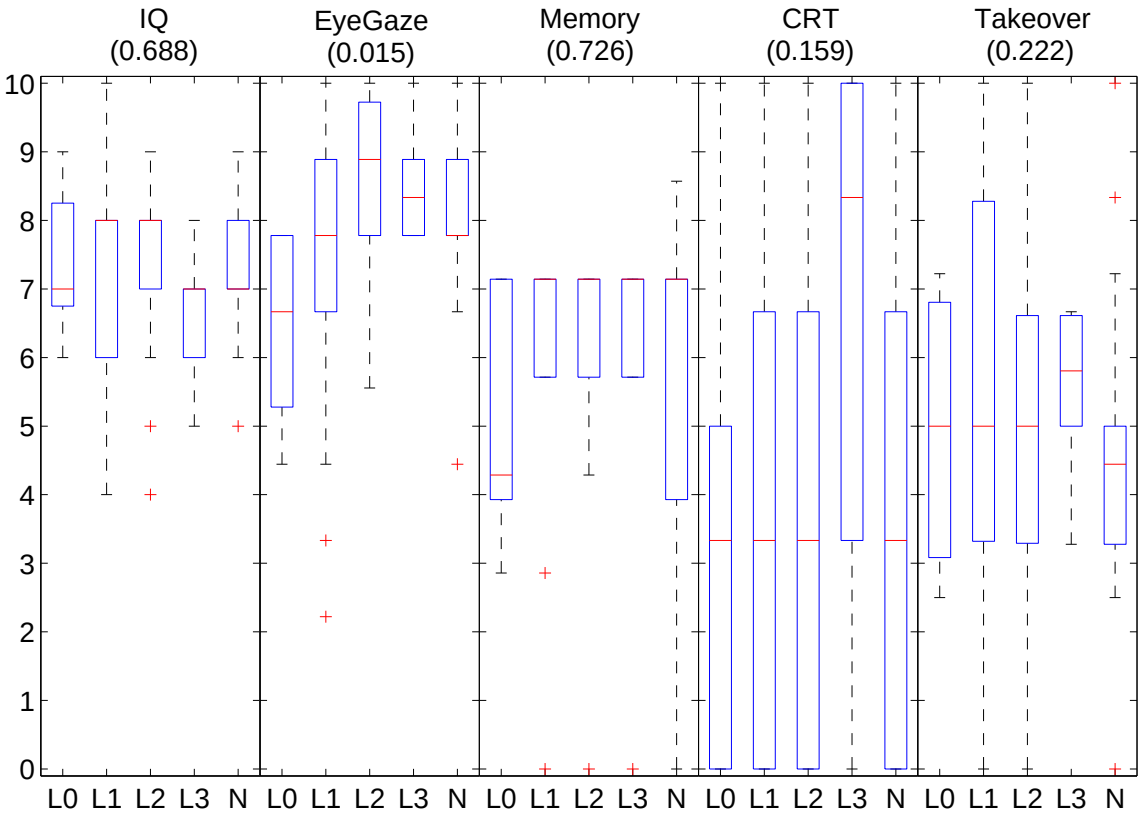

FIGURE IX. Box plots of quiz scores for each estimated level in the pooled undercutting games. p-values in parentheses are for Kruskal-Walis tests that all levels generate the same distribution of quiz scores.

to be Level-1 players. This is clear in Figure VIII, regardless of whether levels are estimated from undercutting games or guess games.

We now ask whether quiz scores can predict estimated levels. We focus here on levels estimated from the pooled families of games; results from individual-game level analyses are qualitatively similar. For each type of quiz, we first perform a Kruskal-Wallis test of the null hypothesis that all five levels' quiz scores are drawn from the same distribution. If this null hypothesis is rejected for some type of quiz, then that quiz can be used to differentiate at least one of the five estimated levels. In that case we perform a mutlinomial logistic regression of levels on that particular quiz score to see which levels have significantly different scores. Since multinomial logistic regressions require an omitted level against which all others are compared, one single regression is not useful

Eye Gaze Score	vs L-0	vs L-1	vs L-2	vs L-3	vs N
Level-0	−−	-0.311	**-0.821**	**-0.865**	**-0.643**
(n=5)		(0.223)	**(0.006)**	**(0.048)**	**(0.020)**
Level-1	0.311	−−	**-0.510**	-0.553	**-0.332**
(n=33)	(0.223)		**(0.011)**	(0.143)	**(0.048)**
Level-2	**0.821**	**0.510**	−−	-0.044	0.178
(n=31)	**(0.006)**	**(0.011)**		(0.909)	(0.361)
Level-3	**0.865**	0.553	0.044	−−	0.222
(n=6)	**(0.048)**	(0.143)	(0.909)		(0.554)
Nash	**0.643**	**0.332**	-0.178	-0.222	−−
(n=41)	**(0.020)**	**(0.048)**	(0.361)	(0.554)	

TABLE IX. Multinomial logit regression coefficient estimates of Eye Gaze quiz scores on pooled undercutting game levels. Each column represents a regression with a different omitted category.

in analyzing all possible comparisons. We therefore report the coefficient estimates from all five possible regressions, where each regression omits a different level.[24]

Figure IX shows a box plot of the distribution of each quiz score for each of the five estimated levels in the pooled undercutting games. The p-values of the Kruskal-Wallis tests for each quiz type appear in parentheses at the top of the graph. We find significant differences across levels only for the Eye Gaze quiz, where Levels 0 and 1 appear to perform worse. The multinomial logistic regression results (Table IX) confirm that Level-0 Eye Gaze scores are significantly lower than those of Levels 2, 3, and Nash, and that Level-1 scores are significantly lower than Level-2 or Nash scores.

The Eye Gaze correlation with Level 0 and 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 low-level actions.

Figure X reports the score distributions for levels estimated from the six pooled guessing games. The Kruskal-Wallis tests indicate that the CRT has some power in predicting subjects' levels. Specifically, the multinomial logistic regressions (Table X) indicate that Level 2 can be distinguished from the two higher levels, but not from the two lower levels.

[24]These five regressions are not meant to be treated as independent tests; rather, reporting them all provides a better view of what is essentially one regression. Using multinomial regression does control for the multiple comparisons within the regression (*i.e.*, within each column).

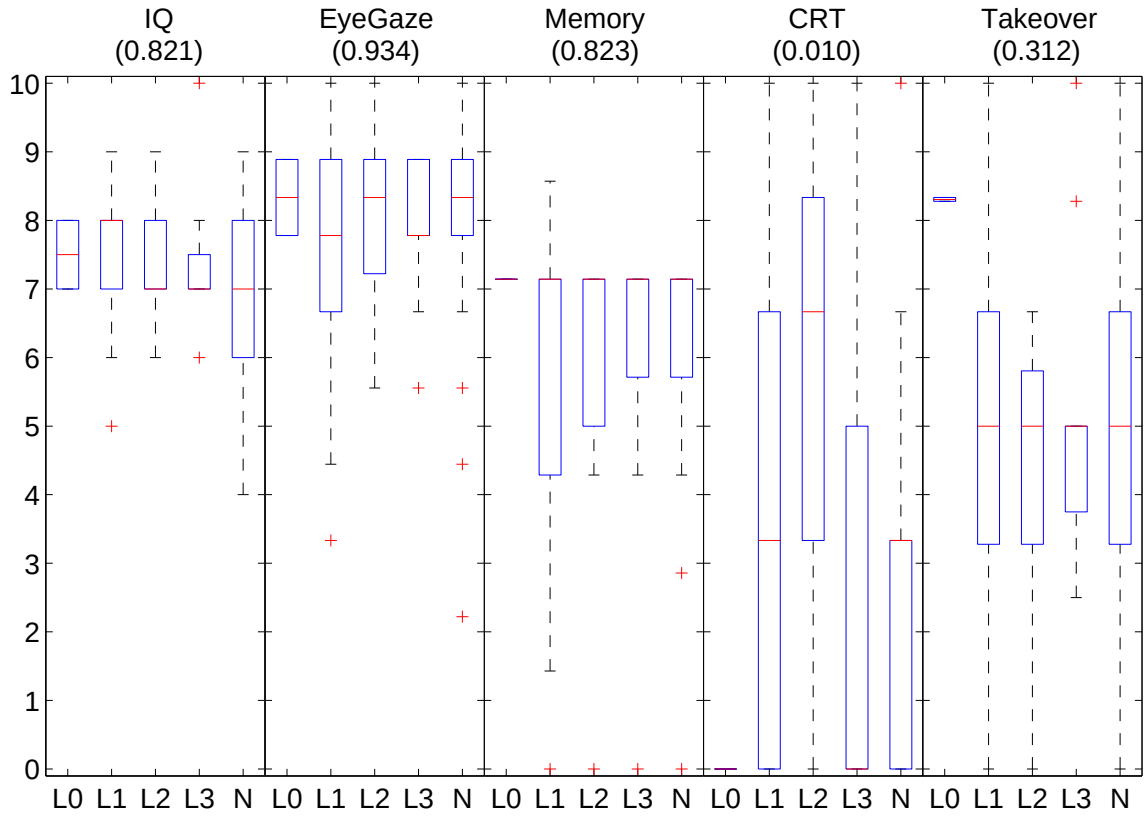

FIGURE X. Box plots of quiz scores for each estimated level in the pooled guessing games. p-values in parentheses are for Kruskal-Walis tests that all levels generate the same distribution of quiz scores.

The correlation between CRT scores and levels is not so transparent, however, given the correlations with expected earnings. The pattern of subjects' expected earnings across levels is hump-shaped (see the right half of Figure VIII). And expected earnings are highly correlated with CRT scores (Table VIII). Thus, it is natural that the pattern of subjects' CRT scores across levels also be hump-shaped.

This three-way correlation allows for multiple interpretations. It may be that the Level-k model i truly descriptive, and CRT scores have a true non-monotonic relationship with strategic sophistication. The linear correlation with expected earnings then comes mechanically from the fact that Level-2 players earn more because the modal subject is a Level-1 type. Alternatively, it may be that the Level-k model is not descriptive, and CRT scores have a true positive relationship with expected earnings. Since most subjects are categorized as Level-1 types (due to its flat likelihood function), subjects

CRT Score	vs L-0	vs L-1	vs L-2	vs L-3	vs N
Level-0	−−	-2.730	-2.883	-2.574	-2.607
(n=2)		(0.824)	(0.815)	(0.834)	(0.832)
Level-1	2.730	−−	-0.153	0.156	0.123
(n=58)	(0.824)		(0.098)	(0.133)	(0.072)
Level-2	2.883	0.153	−−	**0.310**	**0.277**
(n=12)	(0.815)	(0.098)		**(0.018)**	**(0.008)**
Level-3	2.574	-0.156	**-0.310**	−−	-0.033
(n=12)	(0.834)	(0.133)	**(0.018)**		(0.766)
Nash	2.607	-0.123	**-0.277**	0.033	−−
(n=32)	(0.832)	(0.072)	**(0.008)**	(0.766)	

TABLE X. Multinomial logit regression coefficient estimates of CRT quiz scores on pooled guessing game levels. Each column represents a regression with a different omitted category.

with the highest CRT scores (who maximize expected earnings) tend to look as though they are best responding to the Level-1 type. They are therefore classified as Level-2. Other subjects with average CRT scores may be randomly classified as Level-3 or Nash types. The resulting relationship between CRT scores and levels is hump-shaped, as is observed.

We perform similar analyses for game-by-game levels, and the results are consistent with the pooled-game results. In the guessing games, players that play the Level-1 action in at least three of four games have lower Eye Gaze scores than Levels 0, 2, and Nash. They also have higher Takeover Game scores than Levels 0 and Nash. In the guessing games none of the quizzes are diagnostic of levels; the Kruskal-Wallis p-value for the CRT is 0.082 (with subjects estimated to be Level-2 in a majority of games scoring the highest), and greater than 0.15 for all other quizzes.

Result 6: 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.[25] In other words,

[25]Many subjects' responses to a debriefing questionnaire confirm this hypothesis.

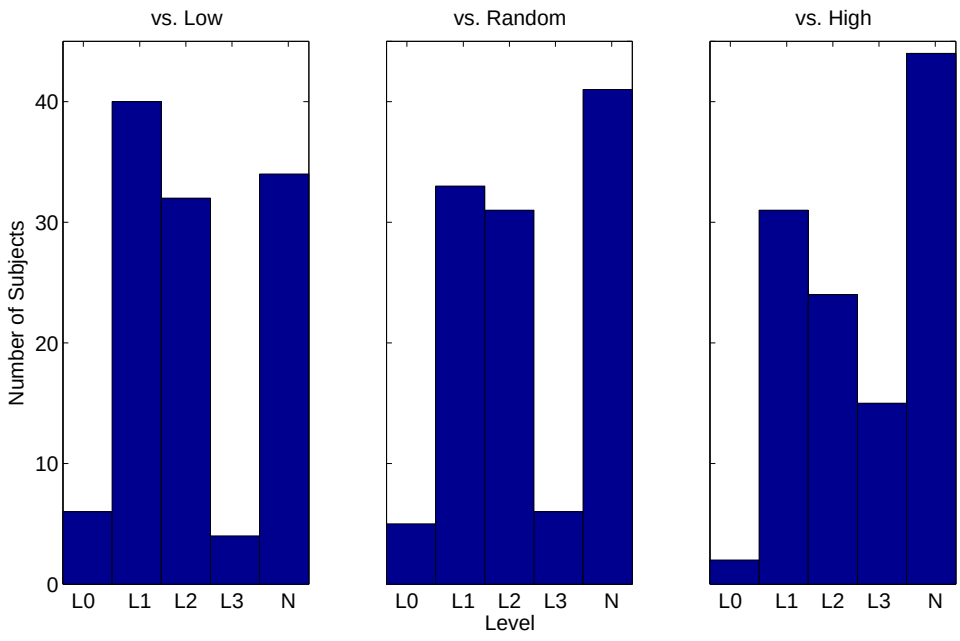

FIGURE XI. Level distributions by opponent in the pooled undercutting games.

information about quiz scores serves as a signal τ_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.

Figure XI shows the histogram of estimated levels in the pooled undercutting games for each of the three opponents. Subjects appear to increase their level of reasoning against stronger opponents. In particular, Level-1 and Level-2 types become less frequent and Nash types more frequent when playing against opponents with higher quiz scores. Kolmogorov-Smirnov (K-S) tests for differences between distributions confirm that the distribution of levels is significantly different between the low-score and high-score opponent (p-value of 0.039), though not significantly different between the low-score and random opponents (p-value of 0.863) or between the random and high-score opponents (p-value of 0.541).

Using game-by-game estimates of levels in the undercutting games gives similar results, with the comparison between low-score and random opponents now being significant (p-value of 0.019).

We next ask whether quizzes predict the magnitude of adjustment. Using the pooled undercutting games, we measure for each subject the difference between their estimated level against a high-scoring opponent and their estimated level against a low-scoring

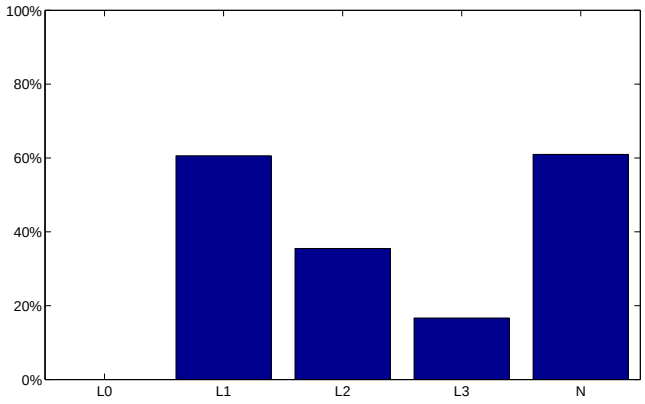

FIGURE XII. For each k, the percentage of Level-k subjects in pooled undercutting games who do not change levels in response to opponent types.

opponent. This difference is then regressed on the five quiz scores. No regression coefficients 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. For each k, we calculate the fraction of Level-k players in the pooled undercutting games whose levels do not shift in response to differing opponents. We refer to these as *stable* players. The percentage of stable players for each k are shown in Figure XII. If players' levels are constrained by their capacities, we should expect that low-level players are more likely to have low capacities, and therefore are more likely to appear as stable players. The data is consistent with this hypothesis for Level-1 through Level-3. Nash types, however, are the most stable. For the capacity-constrained Level-k model to hold, it must be that these players all have high enough capacities so that their chosen level is always greater than four. Such high levels are rarely observed in the literature, suggesting that these players are more likely 'stubborn Nash' types who play Nash equilibrium strategies regardless of the opponent.[26] Thus, there may exist heterogeneity amongst players in the heuristics that they apply.

Similar analyses in the pooled guessing games (Figure XIII) yields no systematic pattern across opponents. None of the three K-S tests for differences show significant differences in distributions (p-values of 0.346, 0.438, and 0.938 for the low-vs-random,

[26]Many authors have included Nash types in heterogeneous behavioral models, but the observation of stubbornness is, to our knowledge, novel.

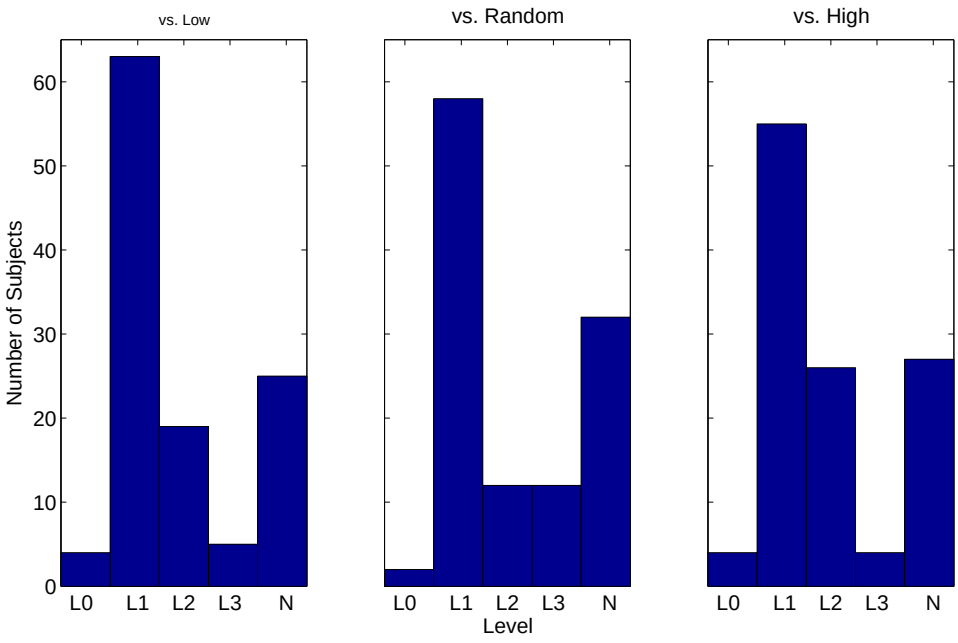

FIGURE XIII. Level distributions by opponent in the pooled guessing games.

random-vs-high, and low-vs-high comparisons, respectively). Using game-by-game estimates yields three virtually identical histograms with K-S test p-values all very close to one.

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.

Result 7: 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 $k_i(\gamma,\tau) \geq k_i(\gamma',\tau)$ for some i and γ then $k_j(\gamma,\tau) \geq k_j(\gamma',\tau)$ for all j. In this way the Level-k model could be thought of as providing a measure of (relative) game difficulty or complexity.

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, or in the opposite direction. These frequencies do not sum to one since pairs where at least one player does not switch levels between games are excluded. The reported frequencies

	Frequency	i.i.d. Prob.
Undercutting Games		
Both change in same direction:	9.6%	27.1%
Both change in opposite directions:	8.5%	27.1%
Opposite/same ratio:	0.884	1.00
Guessing Games		
Both change in same direction:	26.6%	23.8%
Both change in opposite directions:	16.5%	23.8%
Opposite/same ratio:	0.618	1.00
Pooled UGs vs. Pooled GGs		
Both change in same direction:	27.0%	24.9%
Both change in opposite directions:	29.1%	24.9%
Opposite/same ratio:	0.929	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.

are compared against the expected frequencies if levels were drawn independently from the empirical distribution of types.

In the undercutting games we find some support for stability of game orderings. It is more likely that players switch levels in the same direction between games, as opposed to in the opposite direction. A Monte Carlo simulation with 1,000 samples show that the ratio of switch directions is not consistent with the null hypothesis of independently-drawn levels, with a p-value of 0.026. Although this result is statistically significant, its usefulness is tempered by the fact that the vast majority of pairs have at least one player maintaining the same level between games. Thus, a fairly large sample of behavior would be needed to rank games based on observed levels. Analyzing the game-by-game directions of shifts indicates that UG3 is 'easier' than the other three undercutting games. This is also evident from the fact that UG3 has substantially more Nash play than the others. The relative ranks of the other three games is ambiguous. Thus, the ability to rank the undercutting games seems to stem entirely from UG3.

Although the ratio of switch directions is lower in the guessing games, we cannot reject the null hypothesis that the ratio of switch directions was generated by independently-drawn levels. This occurs because the level distributions vary more across guessing games, so the variance of switch directions under the null hypothesis is much larger. A 1,000-sample Monte Carlo simulation yields a p-value of 0.070.

The same techniques can be used to rank the pooled undercutting games against the pooled guessing games. Here, the ratio of switch directions is 0.929, which is not significantly different from one under the null hypothesis (Monte Carlo p-value of 0.380). Thus, the two families of games cannot be clearly ranked using estimated levels.

VIII COMPARISON WITH CGC06

The two-person guessing games used in our experiment were taken from Costa-Gomes and Crawford (2006). In this section we compare our results to their 'Baseline' and 'Open Boxes' data to identify any significant differences. We first use our maximum-likelihood procedure on their raw data to generate levels for each subject in each game, and then repeat all of the above analyses on those levels. Unlike CGC06, we allow for Level-0 types (which account for 9.06% of our data and 9.16% of theirs) but exclude dominance and sophisticated types (which occur in 9.09% of their data).[27]

To our knowledge, there are two notable differences between our experimental design and theirs.[28] First, CGC06 ran their experiments using students from University of California, San Diego and University of York who were enrolled in quantitative courses but did not have extensive training in game theory. Our subjects were taken from a pool of Ohio State University undergraduate students, many of whom are economics majors. We did not select or filter subjects based on their major or courses. In neither experiment did a subject participate in more than one session. Both subject pools appear to be standard within the experimental economics literature.

Second, and perhaps most importantly, the instructions and pre-experiment procedures were substantially different between the experiments. CGC06's subjects read through 19 screens of instructions that included a four-question 'understanding test' in which subjects were required to calculate best-response strategies to hypothetical choices of their opponent, as well as their opponent's best-response strategies to their own hypothetical choices. For example, subjects were asked: "If s/he guesses 500, which of your guesses earns you the most points?", and "If you guess 400, which of her/his

[27]As a robustness check, we use our program to estimate a single level for each subject across all 16 games, as in CGC06, and verify that our level estimates match theirs for every subject, excluding those levels and types that are not common between the two studies.

[28]Certainly other small factors may have influenced behavior, including the exact layout of information on the computer screen, the disposition of the experimenter proctoring the experiment, the room in which the experiment was run, *et cetera*. Given the similarities between the way information was presented on the computer, we expect these to have only second-order effects on behavior, we have no prior expectation about the direction of any such effects, and, in most cases, we have no measure of these differences between experiments.

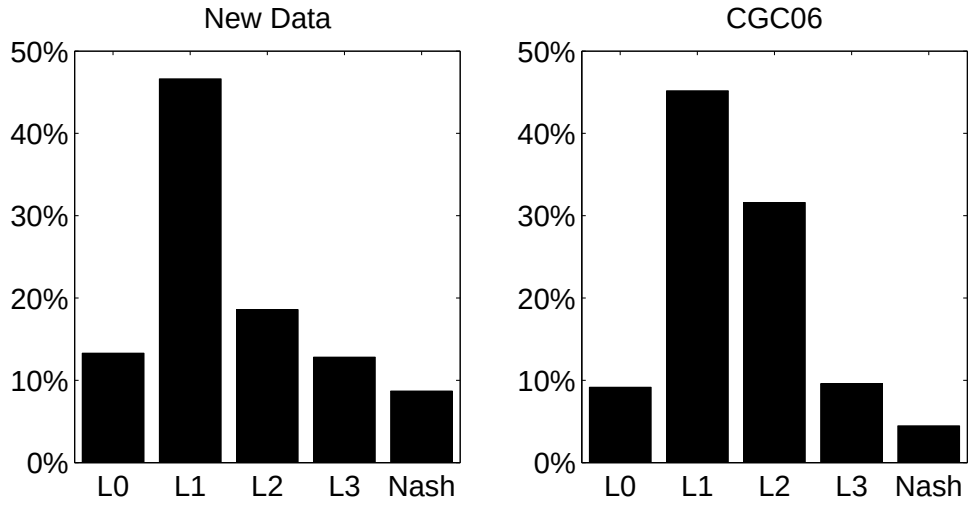

FIGURE XIV. Aggregate level distributions across all standard guessing games in our data and in CGC06.

guesses earns her/him the most points?". Any subject who failed to answer the four questions correctly was not allowed to participate in the experiment.

Our instructions consisted of five printed pages and only informed subjects of how their payoffs are calculated. We did not explicitly ask subjects to calculate best responses (nor opponents' best responses) and we required no test of understanding before proceeding to the experiment.[29] Given the relatively similar subject pools, we expect any differences in behavior between these studies to stem mainly from the instructions and the best-response understanding test.

As seen in Figure XIV, our aggregate distribution of levels among guessing games (estimated game-by-game) looks similar to CGC06, though with fewer Level-2 subjects. Both distributions look typical among the existing Level-k literature. But, as in our data, the game-by-game histograms of levels features a large degree of heterogeneity (Table XII). In games 2–8, for example, we see no Nash types, while in game 13 over 20% of the observations are consistent with the Nash type. Level-1 play varies from 21.59% (game 2) to 73.86% (game 5). These 16 games are ordered by their 'structure'—as defined by CGC06—where lower-numbered games require fewer rounds of dominance elimination to solve the equilibrium, among other considerations. None of the five levels'

[29]We chose this to simplify procedures and make the instructions more concise. Our assessment was that best responses are sufficiently transparent in the guessing games, since minimizing guess error also maximizes payoffs.

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 XII. Frequency of estimated levels in each game of CGC06.

Data	L1	L2	L3	Nash	Total
New Data	14.66%	9.23%	0%	61.67%	14.22%
CGC06	25.36%	21.16%	13.71%	17.61%	19.46%

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

frequencies have a significant correlation with the game number (at the 5% level), indicating that this underlying structure is not driving the variation in level distributions across games.

One of the largest and most obvious differences between CGC06's data and ours is the frequency with which subjects choose strategies that exactly correspond to one of the levels' predictions (excluding Level-0). Only 14.22% of observations correspond to an 'exact hit' in our data, and nearly 20% of CGC06 observations are exact hits (Table XIII). Over 25% of Level-1 observations in the CGC06 data are exact hits, as are over 20% of the Level-2 observations. Our Nash players conform exactly with the predicted strategy more frequently than in CGC06, though the total number of Nash types is relatively low. We believe the differences in exact hit frequencies—especially among Levels 1 and 2—is most likely driven by the difference in instructions between studies and CGC06's use

From To	L0	L1	L2	L3	Nash
L0	20.9%	**44.0%**	23.6%	7.2%	4.4%
L1	8.9%	**54.4%**	24.9%	8.1%	3.6%
L2	6.8%	35.6%	**43.1%**	10.2%	4.2%
L3	6.9%	**38.4%**	33.6%	14.3%	6.9%
Nash	9.0%	**36.5%**	29.8%	14.7%	9.9%
Overall	9.2%	**45.2%**	31.6%	9.6%	4.5%

TABLE XIV. Markov switching matrix of levels in the CGC06 data.

	Frequency	i.i.d. Prob.
CGC06 Guessing Games		
Switch Frequency:	14.8%	22.9%
Non-Switch Frequency:	26.7%	22.9%
Switch Ratio:	0.553	1.00

TABLE XV. Observed frequency of level-switching among pairs of subjects between randomly-drawn games in CGC06's data, compared to the expected frequency under independently-drawn (i.i.d.) types.

of a best-response understanding test, either of which may trigger usage of the Level-k heuristic in subjects.

Surprisingly, the higher frequency of exact hits in the CGC06 data does not translate into noticeably more stable absolute levels. The Markov transition matrix is shown in Table XIV. As in our data, Level-1 acts as an absorbing state, where all subjects have a high probability of transitioning to Level-1, regardless of their current level. The CLPA (constant-level prediction accuracy) of this Markov matrix is 41.9%, so that predicting a player's level in one game based on their observed level in another is accurate 41.9% of the time. Monte Carlo simulations reveal that this is significantly higher (at the 1% significance level) than the 32.3% CLPA expected if individual levels were independently drawn from the population distribution of levels in each game. In absolute terms, a 41.9% CLPA lies between the 34.7% CLPA observed in our guessing games and the 57.6% CLPA in our undercutting games.

The stability of relative levels in CGC06's data also lies between that of our guessing games and our undercutting games. Table XV reveals a switching ratio of 0.553, which lies between the ratio of 0.29 found in our undercutting games and 0.89 in our guessing games. Monte Carlo simulations easily confirm that a switching ratio of 0.553 is not generated by random data (p-value less than 0.001), though it implies that one out of every three pairs of subjects with well-ordered levels will generate a strict switch in their levels between games.

Finally, using levels to order games also generates a result between our guessing game and undercutting game results: The ratio of strict game-order switches over strict non-switches for randomly-drawn pairs of subjects is 0.683, in between the ratio of 0.618 in our guessing games and 0.884 in our undercutting games.

In total, it appears that lengthier instructions and a best-response understanding test creates more stability for the Level-k model, though CGC06's data still reveals considerable instability in level classifications. By almost all measures, the out-of-sample predictive power of the model is better in CGC06's data, though the family of undercutting games provides even higher predictive power (within that family of games), even without longer instructions and an understanding test.

Crawford et al. (2010) argue that a best-response understanding test is crucial for replicating field settings because "most people seem to understand very well how their payoffs are determined" (p. 32). Although we did not require an understanding test, our instructions provided adequate and simple descriptions of subject payoffs. For example, subjects in our experiments were told "you will be paid for this game based on how small your error is, and smaller errors mean larger payoffs", mathematical formulas for calculating errors and payoffs were given along with verbal descriptions, payoffs (as a function of errors) were shown in graphical form, and two numerical examples were worked out. In a post-experiment questionnaire, we received no feedback that subjects were confused about payoffs in any of the games.

We view differences between these studies as evidence that the Level-k model's predictions are not robust to varying protocols, as varying the instructions and understanding tests may trigger different behavioral heuristics within the same game. Applying any one behavioral model to the field may require some attention to the level of instruction or training that agents have received. Unfortunately, these factors may be difficult to quantify, heterogeneous across agents, or unobservable. Uncertainty about past experiences would then lead to uncertainty about the predictive accuracy of the Level-k model.

IX Discussion

In sum, our results are mixed: The Level-k model exhibits reasonably strong cross-game stability within the family of undercutting games, but virtually none in the two-person guessing games. Even in the undercutting games, however, observed levels do not correlate well with our five psychometric measures, except that Level-1 players may have shorter working memory and a less keen awareness of others' emotions and cognition. Finally, it appears that some players 'step up' against stronger opponents, but we are

unable to predict who makes this adjustment using either psychometric measures or observed levels.

Although ours is the first paper to thoroughly examine cross-game stability of individual levels, our conclusions about the success of the Level-k model are broadly consistent with the past literature. Many papers find strong support for Level-k play in certain games using behavioral data alone (Stahl and Wilson, 1994, 1995; Nagel, 1995; Duffy and Nagel, 1997; Ho et al., 1998; Bosch-Domènech et al., 2002; Camerer et al., 2004) or behavioral data augmented with lookup data (Costa-Gomes et al., 2001; Costa-Gomes and Crawford, 2006) or eye-tracking data (Chen et al., 2009; Wang et al., 2009). For some games, however, the Level-k model does not appear to organize the data well (Ivanov et al., 2009b,a; Crawford and Iriberri, 2007b).[30] Shapiro et al. (2009) even find that the model's fit can vary within a single game when different components of the payoff function are emphasized, with a better fit as the game becomes closer to a standard p-beauty contest and a worse fit as the game approaches the incomplete-information global game of Morris and Shin (2002). The broad conclusion that emerges from this line of research is that the Level-k approach works well in some games, perhaps accurately describing how subjects make choices, while working less well in other games.

Camerer et al. (2004, p. 873) argue that "fitting a wide range of games turns up clues about where models fail and how to improve them." Our research represents one such contribution, by demonstrating the varying individual-level robustness of Level-k models across two families of games. Thus, the Level-k model may be one of many heuristics players use to select strategies, with different heuristics being triggered unconsciously in different settings. Beauty contests, simple matrix games, and our undercutting games all seem to trigger the Level-k heuristic in a large fraction of subjects, while its use appears infrequent in common-value auctions, global games, and endogenous-timing investment games. In two-person guessing games the Level-k heuristic does not appear to be triggered unless subjects are trained to calculate iterated best responses prior to play.

In its full generality, the triggered-heuristics meta-model is not falsifiable, since each game in each situation could have a unique heuristic. But applying data to the model helps restrict its parameters by identifying regularities across games. The Level-k model has already been identified as one heuristic that is likely used in many games and in many situations. Our results show that it is a persistent and robust heuristic

[30] Crawford and Iriberri (2007b) point out that the Level-k model fails to account for overbidding in second-price auctions.

across one family of games. Understanding the boundaries of the domain of applicability of the Level-k model means understanding when it is used, when it is not, and what factors trigger its use; this, in turn, increases the overall predictive power.

At this point, we conjecture that Level-k play is triggered by simple, normal-form games of complete information, as well as in situations where the game's instructions directly focus attention on calculating best responses, either directly through understanding tests or indirectly through framing effects. In games of incomplete information and games that are not represented in matrix form we expect alternative heuristics to prevail. These hypotheses give rise to a wide range of open questions that can be addressed in future work.

Given our conclusions, we suggest focusing behavioral research both on identifying new heuristics *and* exploring their triggers. For example, Ivanov et al. (2009b) identify plausible 'rules of thumb' to explain their data when Level-k and quantal response equilibrium cannot; to what extent do these heuristics extend beyond the dynamic investment game they study? In our two-person guessing games, we do not identify an alternative heuristic that organizes the data since our analysis focuses on players' estimated levels and not their actual strategies; however, our results make strides in the right direction. For example, strategies do not seem to depend critically on the best response function of the game (or even logistic best response) because aggregate play is unchanged between the standard and zero-sum versions of the game, and also because subjects who are trained to calculate iterated best responses behave differently from those who are not. But what triggers this particular non-best-response heuristic remains unexplored.

APPENDIX A ROBUSTNESS TO THE NUMBER OF GAMES PER ESTIMATE

In this appendix we briefly explore the robustness of Level-k estimates to the number of games used in each estimate.[31] It may be that assigning a single level to each observation introduces significant noise in the resulting levels, causing the results to appear artificially biased toward randomly-generated levels. Estimating levels based on multiple games may reduce this variability and lead to more reliable estimates of players' types, leading to greater stability in the Level-k model.

Formally, let $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_m\}$ denote the set of m games played by the subjects. For each divisor r of m one can construct partitions of the form $P_{m,r} = (p_1, \ldots, p_r)$ of Γ consisting of r sets of m/r games each. For example, if $m = 6$ and $r = 3$ then one possible

[31] We thank Vince Crawford for suggesting this test.

partition of the 6 games into 3 sets is $P_{6,3} = \{\{1,2\},\{3,4\},\{5,6\}\}$. Letting $s = m/r$, the number of partitions containing r equal-sized sets of s elements each is given by

$$q(m,s) = \frac{\binom{m}{s}\binom{m-s}{s}\binom{m-2s}{s}\cdots\binom{s}{s}}{\frac{m}{s}!}.$$

Note that $q(m,m) = q(m,1) = 1$. Let q index the various partitions from 1 to $q(m,s)$, so $P^q_{m,r}$ is one of the partitions of m games into r equal-sized subsets.

Take any set of data from n players over m games, and any divisor r of m. We can pick any $q \in \{1,\ldots,q(m,r)\}$, take the partition $P^q_{m,r} = \{p_1, p_2, \ldots, p_r\}$, and for each partition element p_j, estimate a level for each subject i over the set of games in p_j. This is done exactly according to the maximum-likelihood procedure used in CGC06 and in this paper, where the likelihood of observing data point x under level k is given by a logistic error structure around the optimal strategy for k, with a 'spike' of weight ε on the exact Level-k strategy. The result is a level estimate for each player i in each partition element p_j, which we denote simply by $k_i(j)$. Thus, we generate r levels for each subject, using m/r games (or, data points) for each level estimated.

In CGC06 r always equals one; in our paper r either equals m (for game-by-game analyses) or one (for pooled analyses). In either case $q(m,s) = 1$, so the choice of which partition to choose is trivial. Here we explore intermediate cases where $1 < r < m$. Ideally, we would fix r, generate all possible partitions of size r, and for each partition, generate r estimated levels per subject. We could then perform analysis of the stability of those r levels (as in the body of the paper). For example, the switch ratio can be calculated for each partition $q \in \{1,\ldots,q(m,s)\}$ and the entire 'distribution' of $q(m,s)$ switch ratios reported.

Since $q(m,s)$ can be quite large ($q(16,4) = 2,627,625$, for example), we instead draw a small random sample of possible partitions. We then estimate r levels per subject, calculate the switching ratio for each randomly-drawn partition, and report the sample distribution of switch ratios. We perform this exercise for each divisor r of m to see how the distribution of switch ratios would change as more games are used per level estimate (or, equivalently, as fewer level estimates per subject are performed). This is done for both our guessing game data (where $m = 6$) and the CGC06 guessing game data (where $m = 16$).

The results of this analysis appear in Figure XV. The horizontal axis contains the various values of r. The case of $r = 1$ is degenerate; each subject has only one level estimate and so stability measures such as the switching ratio are not defined. The vertical axis reports the switching ratio, as described in the body of the paper.

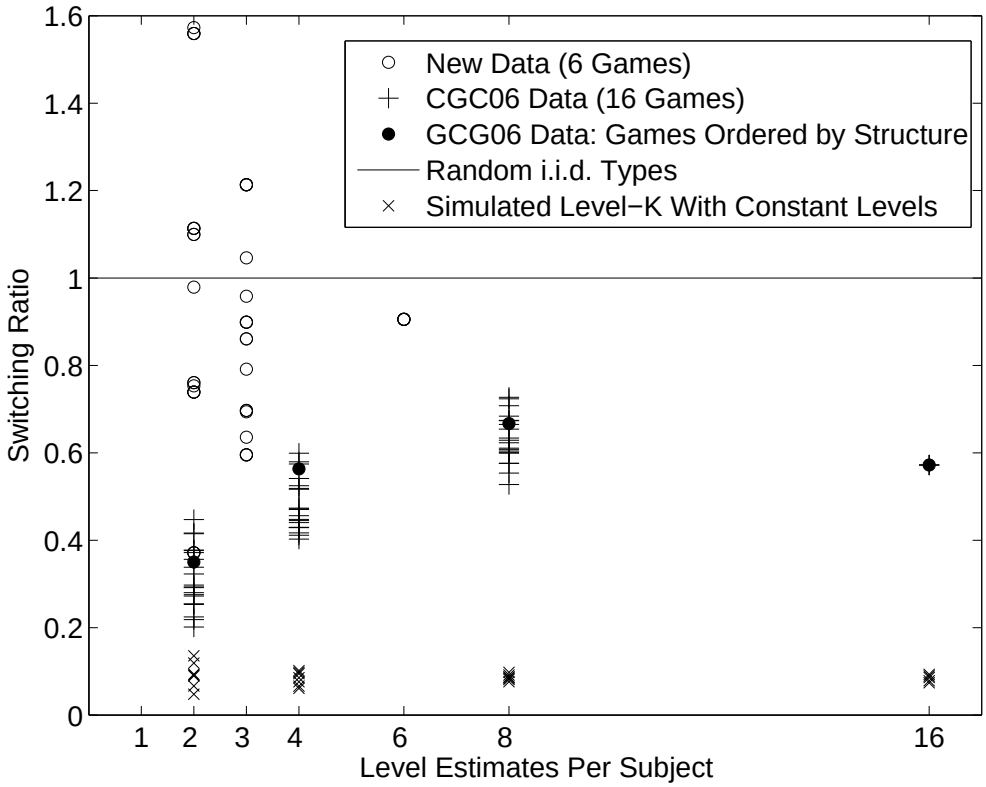

FIGURE XV. Switching ratios as the number of levels estimated per subject varies, using many randomly-drawn partitions of the games.

As benchmarks, we include a horizontal line at one to indicate the switching ratio if the levels were independent random draws from a fixed distribution. We also simulate the switching ratio for the Level-k model with constant k_i functions; in theory these ratios should all equal zero, but because a true Level-0 subject (who randomly selects their strategy) would occasionally be misclassified as a different level, some randomness is introduced into the level estimates. This can result in a small but non-trivial switching ratio.

As the number of estimates per subject decreases, so too does the frequency with which randomly-drawn subjects can be strictly ordered by their levels in two randomly-drawn games. Thus, both the numerator and denominator of the switching ratio become smaller as r decreases; this generates higher variance in the switching ratio distributions for small r.

CGC06 order their games based on 'structure', roughly corresponding to how many steps of elimination of dominated strategies are necessary to solve the Nash equilibrium of the game. We report the switching ratios for the partitions that respect this ordering. Specifically, if $\{1,2,\ldots,16\}$ is the original ordering of the 16 games, we report the switching ratios for the partitions $\{\{1,\ldots,8\},\{9,\ldots,16\}\}$, $\{\{1,\ldots,4\},\ldots,\{13,\ldots,16\}\}$, $\{\{1,2\},\{3,4\},\ldots,\{15,16\}\}$, and $\{\{1\},\{2\},\ldots,\{16\}\}$.

The graph reveals that stability in the CGC06 data improves with fewer estimates per subject (or, more games per estimate), though its switching ratios never overlap with the constant-level switching ratios. In the best case ($r = 2$) the switching ratios approach the 0.288 ratio achieved in our undercutting games. The ordering of CGC06's games based on structure, however, does not generate obviously greater or smaller switching ratios. Switching ratios in our data do not improve with more games per estimate. This suggests that CGC06's subjects were somewhere more persistent in their underlying type and in fact there was some noise added to their estimated levels by using only one game per estimated (or, more correctly, assigned) level.

Again, the most obvious difference in experimental design between CGC06 and our experiment is in the length of instructions and use of an understanding test. We therefore speculate that one or both of these design features triggered the use of the Level-k heuristic in more subjects in the CGC06 experiment than in ours. This results in relatively more stable level estimates across games for their data.

REFERENCES

Altmann, S., Falk, A., 2009. The impact of cooperation defaults on voluntary contributions to public goods, univeristy 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-blindnessŠ. 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.

Bhatt, M., Camerer, C. F., 2005. Self-referential thinking and equilibrium as states of mind in games: fmri evidence. Games and Economic Behavior 52 (2), 424–459.

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.

Brocas, I., Camerer, C., Carrillo, J. D., Wang, S. W., 2009. Measuring attention and strategic behavior in games with private information, cEPR Discussion Paper No. DP7529.

Brown, A. L., Camerer, C. F., Lovallo, D., 2010z. To review or not to review? limited strategic thinking at the movie box office, working paper.

Bruguier, A. J., Quartz, S. R., Bossaerts, P. L., 2008. Exploring the nature of "trading intuition", california Institute of Technology working paper.

Burnham, T. C., Cesarini, D., Johannesson, M., Lichtenstein, P., Wallace, B., 2009. Higher cognitive ability is associated with lower entries in a p-beauty contest. Journal of Economic Behavior and Organization 72, 171–175.

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.

Chong, J.-K., Camerer, C. F., Ho, T.-H., 2005. Cognitive hierarchy: A limited thinking theory in games. In: Zwick, R., Rapoport, A. (Eds.), Experimental Business Research: Marketing, Accounting and Cognitive Perspectives. Vol. III. Springer, The Netherlands, Ch. 9, pp. 203–228.

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., Costa-Gomes, M. A., Iriberri, N., December 2010. Strategic thinking, oxford University working paper.

Crawford, V. P., Iriberri, N., 2007a. Fatal attraction: Salience, naivete, and sophistication in experimental "hide-and-seek" 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.

Gigerenzer, G., 2001. The adaptive toolbox. In: Gigerenzer, G., Selten, R. (Eds.), Bounded Rationality: The Adaptive Toolbox. MIT Press, London, Ch. 3.

Grosskopf, B., Nagel, R., 2008. The two-person beauty contest. Games and Economic Behavior 62, 93–99.

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., 2009a. Can relaxation of beliefs rationalize the winner's curse?: An experimental study, forthcoming in *Econometrica*.

Ivanov, A., Levin, D., Peck, J., 2009b. Hindsight, foresight, and insight: An experimental study of a small-market investment game with common and private values. American Economic Review 99 (4), 1484–1507.

Kawagoe, T., Takizawa, H., 2009. Equilibrium refinement vs. level-k analysis: An experimental study of cheap-talk games with private information. Games and Economic Behavior 66 (1), 238–255.

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.

Morris, S., Shin, H. S., 2002. Social value of public information. American Economic Review 92, 1521–1534.

Nagel, R. C., 1993. Experimental results on interactive competitive guessing, discussion Paper 8-236, Sonderforschungsbereich 303, Universitat Bonn.

Nagel, R. C., 1995. Unraveling in guessing games: An experimental study. American Economic Review 85 (5), 1313–1326.

Ostling, R., Wang, J. T.-y., Chou, E., Camerer, C. F., 2009. Testing game theory in the field: Swedish lupi lottery games, working paper.

Rogers, B. W., Palfrey, T. R., Camerer, C. F., 2009. Heterogeneous quantal response equilibrium and cognitive heirarchies. 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.

Shapiro, D., Shi, X., Zillante, A., October 2009. Robustness of level-k resaoning in generalized beauty contest games, university of North Carolina working paper.

Siegel, S., Castellan, Jr., N. J., 1988. Nonparametric Statistics for the Behavioral Sciences, 2nd Edition. McGraw-Hill, New York, NY.

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.

Wang, J. T.-y., Spezio, M., Camerer, C. F., 2009. Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. American Economic Review 100 (3), 984–1007.

Wechsler, D., 1958. The measurement and appraisal of adult intelligence, 4th Edition. Williams & Wilkins, Oxford, England.