On the Persistence of Strategic Sophistication

Sotiris Georganas[a], Paul J. Healy[b*] and Roberto A. Weber[c]

[a]*Dept. of Economics, City University London, Northampton Square, London, EC1V 0HB, U.K.; sotiris.georganas.1@city.ac.uk.*
[b]*Dept. of Economics, The Ohio State University, 1945 North High street, Columbus, OH 43210, U.S.A.; healy.52@osu.edu.*
[c]*Dept. of Economics, University of Zurich, Blumlisalpstrasse 10, 8006 Zurich, Switzerland; roberto.weber@econ.uzh.ch.*

ABSTRACT. We examine whether the "Level-k" model of strategic behavior generates reliable cross-game predictions at the individual level. We find no correlation in subjects' estimated levels of reasoning across two families of games. Furthermore, estimating a higher level for Ann than Bob in one family of games does not predict their ranking in the other. Direct tests of strategic reasoning generally do not predict estimated levels. Within families of games, we find that levels are fairly consistent within one family, but not the other. Our results suggest that the use of Level-k reasoning varies by game, making prediction difficult.

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

JEL Classification: C72; C91; D03.

[*] Corresponding author.

1 INTRODUCTION

Following a considerable literature demonstrating deviations from Nash equilibrium play (see, for example, 16), 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 [53, 62, 24, 18].[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., 18), 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. Estimated levels in one game could then be used to predict players' behavior in novel games. Moreover, estimates of a player's level could be improved by using direct psychometric measures that correlate with strategic sophistication.

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 the model is accurate but players' levels vary

[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 [50]. [57] 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 25, 32, 17, for example].

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. 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 [23] (henceforth CGC06).[3]

Within each family of games, we identify an individual's level in a Level-k framework, following a standard approach for classifying individual behavior based on the observation of play in several games. We then test whether these observed levels satisfy any of the cross-game restrictions we have identified. To complement this analysis, we also attempt to identify individual levels separately for each game, and use these classifications to conduct cross-game comparisons within each family of games.

We also consider two additional ways in which strategic sophistication might be detectable. First, we elicit several direct measures of strategic intelligence using brief quizzes that have been found to identify strategic 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:

[2] We do not suggest that levels must be constant across games for the model to have predictive power. [18] and [20], for example, suggest that levels *will* change in certain situations. Predictive power simply requires that situational changes be predictable.

[3] A two-person guessing game is different from the two-person beauty contest studied by [38]; the latter has a (weakly) dominant strategy while the former does not.

(1) The *aggregate* distribution of levels is similar to that found in previous studies for both families of games.
(2) Individual levels show little persistence between the two families of games. Moreover, the relative ordering of players is also unstable between the two game families.
(3) Looking within families of games, the aggregate distribution of levels is remarkably stable across undercutting games, but quite unstable across two-person guessing games. Individual levels and relative ordering are moderately persistent within the family of undercutting games, but have no persistence within the family of guessing games.
(4) The quizzes generally fail to predict players' levels in either family of games, though Level-1 play is correlated with a test for autism in undercutting games.
(5) Some players adjust strategies against stronger opponents, but neither quiz scores nor levels predict which subjects make this adjustment.

Our interpretation of these results is that the congruence between Level-k models and subjects' actual decision processes depends on the context. Players confronted with a novel game may have many alternative processes for determining what strategy to select, and different environments trigger the use of different decision processes.[4] Level-k reasoning might be one process that is triggered in some contexts (undercutting games), but not in others (guessing games). Of course, if Level-k reasoning is employed in some games, it is critical for the theory's predictive ability to be able to identify what factors trigger its use.

Additional insight into robustness comes from comparing our guessing game data with that of CGC06.[5] In their data the Level-k model receives stronger support—especially when considering subjects' "lookup" behavior—though we also find substantial cross-game instability in their data as well. We believe this difference in model accuracy stems from two differences in protocols. First, their instructions are far more detailed (spanning 31 computer screens) and include four practice rounds with feedback on aggregate choices. Ours consist of a 5-page handout with no practice rounds. Second, they require that subjects pass an understanding test in which they must calculate their best response to an opponent's choice, and their opponent's best response to their own choice. Subjects who fail this test are dismissed. We have no such understanding test, and include all subjects. It is possible that the more extensive instructions, the practice rounds with feedback, and the best-response understanding test all trigger the use of Level-k reasoning in a greater fraction of the subjects. Alternatively, it is possible that our

[4] This interpretation is similar to the idea of a "toolbox" of various decision making approaches or heuristics, which are employed varyingly depending on the context [36].

[5] Our data are comparable to the CGC06 "Baseline" and "Open-Box" treatments. See the online appendix for this and other comparisons with CGC06.

design fails to eliminate confusion sufficiently, leading to noisier data. Regardless, the sensitivity of results to the protocol suggests that the Level-k model's applicability may be limited in this regard.

2 Review of Relevant Literature

The notion of heterogeneous strategic sophistication operating through limited iterations of best response dates back at least to the "beauty contest" discussion of [47]. Motivated by this, Nagel [53, 54], Ho et al. [42], and others study behavior in laboratory p-beauty contest games, in which all players submit a number in $[0, 100]$ and the closest guess to p times the average wins a prize. The observed distributions of guesses show clear spikes consistent with Level-1 and Level-2 play. This finding is robust to the structure of the game [31] and varied populations [10].[6]

Stahl & Wilson [62] study Level-k behavior in ten 3×3 matrix games.[7] 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 [63] examine play in twelve normal-form games played without feedback, adding Worldly and Rational Expectations types. In both studies, many subjects fit strongly into one type, with posterior probabilities of their maximum likelihood type exceeding 0.90. Stahl & Wilson Stahl and Wilson [63] also provide 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 percent. Using this relatively low threshold, they find that 35 of 48 subjects are stable. In contrast, we estimate a player's type independently in two sets of games, and directly compare whether the two estimated types are identical.[8]

Costa-Gomes, Crawford & Broseta Costa-Gomes et al. [24] fit a Level-k model, with 9 possible types, to behavior in 18 matrix games. In their experiment, payoffs in the games are initially hidden to subjects, so that estimation of a player's level based on strategy choice

[6] When the game is made into a global game in which players are also rewarded for guessing an unknown state [52], however, the Level-k model fits poorly when the state-guessing incentive is emphasized [60].

[7] 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.

[8] Burchardi & Penczynski [14] and Penczynski [56] find that players' estimated levels are altered after communicating with others. Although this represents one notion of type instability, it is unlikely that it stems from true randomness in players' types.

can be augmented by analyzing which pieces of information subjects choose to view before making a decision. The model fits well, and they generally find higher levels in simpler games.

Camerer, Ho & Chong [18] introduce the *Cognitive Hierarchy* variation of the Level-k model, in which players best respond to the 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.

In Costa-Gomes & Crawford Costa-Gomes and Crawford [23] (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. As in the earlier paper Costa-Gomes et al. [24], lookup behavior is used to strengthen type estimation. 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 our analysis. Chen, Huang & Wang [19] 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.

Arad & Rubinstein [2] introduce the 11-20 money request game, which is similar to our undercutting game in that it is a simple game designed to trigger Level-k behavior while allowing a clean separation of levels. Although they are not explicitly testing for cross-game stability, they do find that subjects behave differently across variations of the game that do not change the equilibrium or Level-k predictions. DeSousa, Hollard & Terracol [29] identify non-strategic players by observing play in 10 beauty contest games, and find that these subjects are more likely to play non-strategically in the 11-20 money request game as well. Arad &

Rubinstein [3] develop a model of multi-dimensional iterative reasoning, focused on ‘features’ of strategies, and apply it to behavior in Colonel Blotto games. They find that subjects who apply more iterative reasoning in an 11-20 money request game also seem to exhibit more multi-dimensional iteartive reasoning in the Colonel Blotto game.

Batzilis *et al.* [8] fit a Level-k model to a very large number of rock-paper-scissors (‘Roshambull’) games among Facebook users. They find that aggregate play frequencies vary slightly from the equilibrium prediction, and that most players’ strategies are not consistently aligned with any single level.

The Level-k model has also been applied to extensive-form games. Kawagoe & Takizawa [46] study centipede games and compare 12 different specifications of the Level-k / Cognitive Hierarchy model against two specifications of the Agent Quantal Response Equilibrium (AQRE) model [51]. They find that an AQRE specification fits best for their constant-sum centipede game, while a Cognitive Hierarchy model with a uniform Level-0 strategy fits best their increasing-sum centipede game. Ho & Su [41] develop a dynamic version of the Level-k model for centipede games and show that it fits both first-round play and the pattern of earlier taking as the game is repeated. The model also matches patterns of behavior in a dynamic bargaining experiment.

Relatively few authors test whether estimated levels correlate with personal traits such as intelligence. Camerer, Ho & Chong [18] find higher average levels in subject pools with greater academic training, such as Caltech undergraduates and game theorists. Burnham *et al.* [15] show that individuals’ choices in a p-beauty contest game correlate with scores on a 20-minute test of cognitive ability. Gill & Prowse [37] also find a correlation between cognitive ability and levels in a p-beauty contest, and show that higher-ability players are more likely to converge toward equilibrium over time and earn higher payments. Chong, Camerer & Ho [20] find that cognitive effort matters along with intelligence. They let their subjects play 22 mixed-equilibrium matrix games in a fixed order and report a positive correlation between thinking time and levels. Furthermore, average levels are higher in games 12-22 than in games 1–11, indicating a learning-by-doing increase in sophistication over time.[9] Rubinstein [58] finds correlation in players’ reaction times across games, suggesting that some systematically engage in more contemplation than others. But he reports that he could not find interesting cross-game correlations in strategy choices, and that the level of contemplativeness is not very predictive

[9] 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. 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.

of strategy choices. Similarly, DeSousa *et al.* [29] do not find a correlation between the Elo ranking (a measure of a chess player's quality) and the likelihood of playing strategically.

The Level-k model has also been applied successfully to a variety of other games, including "hide-and-seek" games [27], incomplete-information betting games [11], betting games and matrix games [57], coordinated attack games [48], sender-receiver games augmented with eye-tracking data [65], and cheap-talk games [45]. In the field, Level-k has been shown to fit behavior in Swedish lowest-unique-positive-integer lottery games [55] and to explain the fact that movies that were not released to critics before their public opening earn higher revenues [12]. A functional MRI study even suggests differences in brain activity between subjects who exhibit varying degrees of "strategic sophistication" [9]. Finally, a few recent papers apply the Level-k concept to study departures from Nash equilibrium play in auctions, finding that the Level-k approach often, though not always, yields a significantly better fit than the Nash equilibrium [26, 35]. However, Ivanov *et al.* [44] show that models with misguided beliefs (such as Level-k) cannot explain the winner's curse in common value auctions, because subjects who play against their own past actions still exhibit substantial overbidding.

For a more comprehensive survey of studies on the Level-k model, see [28].

3 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, or test only how the aggregate distribution of types varies across games or populations [18, e.g.]. In this section we introduce a formal framework in which such testable restrictions can be defined clearly. Our experiment then examines 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.[10]

[10] For example, [24], [23], [26], and [27] assume that all players with a level of $k > 0$ believe all other players' level to be $k-1$ with probability one. [18], on the other hand, assume that all players with a level of $k > 0$ believe the realized levels of 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

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 denote the signal i receives 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.[11] 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 $k_i \in \mathbb{N}_0$.[12] For example, in [18], $\lambda > 0$ 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 response to beliefs $\nu(k)$, given that each level $\kappa < k$ of player j plays σ_j^κ.[13] Formally, for each $k > 0$,

(i believes j believes his opponent's levels follow this distribution, *et cetera*). Strzalecki [64] 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.

[11] 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.

[12] 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.

[13] 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.

the strategy σ_i^k 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.[14]

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 $\{\sigma_i^0, \sigma_i^1, \sigma_i^2, \ldots\} \cup \{\sigma_i^N\}$.[15] 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.

Examples of possible restrictions on k_i that one can test using 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.

[14] 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.

[15] 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.

(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, but to otherwise keep levels constant across games.

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.[16]

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 a given level 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 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 some identifiable restrictions.

[16] Cross-game learning may still generate violations 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 1. 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 2. Undercutting game 2 (UG2).

4 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 [23].

4.1 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, $s_i \in \{1, 2, \ldots, m, \ldots, n\}$. 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 her best response is to choose $s_i = m$; otherwise her best response is to "undercut" her opponent by choosing $s_i = s_j - 1$. If no player undercuts the other then one of the following situations apply: If both choose $s_i = 1$ (the

	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 3. 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 4. Undercutting game 4 (UG4).

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 to rule out any mixed-strategy Nash equilibria.

The payoff matrices of the undercutting games used in this experiment are shown in Figures 1–4. Consider UG1, shown in Figure 1. A levels-of-reasoning model that assumes uniformly random play by Level-0 types will predict that Level-1 types play $s^1 = 4$ as it maximizes the sum of row payoffs, Level-2 types play $s^2 = 3$, Level-3 types play $s^3 = 2$, and all higher levels play the equilibrium strategy of $s^N = 1$. 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. Moreover, variations in the assumed Level-0 strategy have no impact on the ordering of players' inferred levels. For example, if i plays 3 and j plays 2 then we infer that $k_j = k_i + 1$, regardless of the Level-0 specification.

4.2 Two-Person Guessing Games

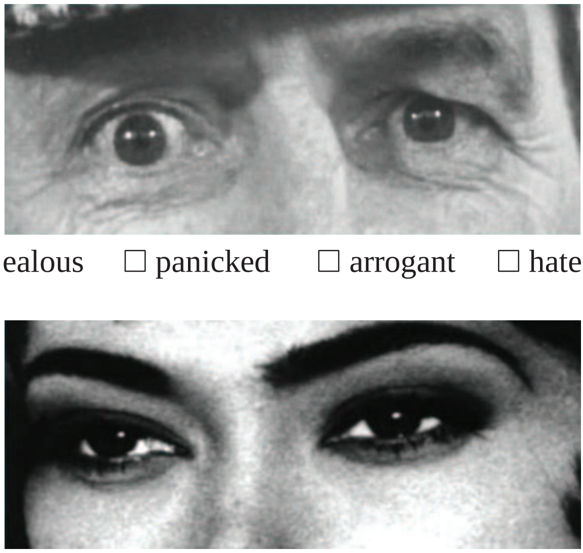
Two-person guessing games are asymmetric, two-player games parameterized by a lower bound $a_i \geq 0$, upper bound $b_i > a_i$, and target $p_i > 0$ for each player. Strategies are given by $s_i \in [a_i, b_i]$ and player i is paid according to how far her choice is from p_i times s_j, denoted by $e_i = |s_i - p_i s_j|$.

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 23].

5 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 completed five tasks, intended to measure general cognitive ability and strategic reasoning:

(1) an IQ test,
(2) the Eye Gaze test for adult autism,
(3) the Wechsler digit span working memory test,



□ jealous □ panicked □ arrogant □ hateful

□ aghast □ fantasizing □ impatient □ alarmed

FIGURE 5. Sample questions from the Eye Gaze 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.[17] Similar tests of cognitive ability have been shown to correlate with higher levels of reasoning in p-beauty contest games [15].

The Eye Gaze test [7] asks subjects to identify the emotions being expressed by a pair of eyes in a photograph. See Figure 5 for sample problems. Poor performance on this task is diagnostic of high-functioning adult autism or Asperger's Syndrome [7] 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 [13].

The Wechsler Digit Span memory test tests subjects' abilities to recall strings of digits of increasing length. It is one component of the Wechsler Adult Intelligence Scale [66] to assess overall intelligence. [30] had 67 subjects take this short-term memory test and then play three games against a computerized opponent that always selected the equilibrium strategy. The three games all required iterated reasoning to solve the equilibrium best response. They found

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

a positive and significant correlation between subjects' memory test score and the frequency with which they selected the best response.[18]

The CRT contains three questions for which the ininitial intuitive 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 [34]. This measure also correlates with a tendency to play default strategies in public goods games [1].

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 [59], even after receiving feedback and gaining experience [4].

We normalized each of the quiz scores to a scale of ten possible points. For scoring purposes during the experiment, we combined the CRT and Takeover game into one four-question, ten-point quiz, with answers coded using a binary (correct or incorrect) classification. For the Takeover game, subjects received a positive score if and only if their bid was exactly zero—the unique profit-maximizing bid.[19] The sum of the four quiz scores was calculated for each player. Players were given no feedback about any player's absolute or relative performance on the quizzes until the end of the experiment, at which point they learned only their own total quiz score.

After completing the quizzes, the subjects played ten games against varying opponents. The first four games are undercutting games and the last six are guessing games.[20] The parameters of each game are given in Table 1 and Figures 1–4. The final three guessing games are identical to the first three, with the players' roles reversed. As in [23], 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 were asked to choose a strategy against a random opponent, against the opponent (other than themselves) with the highest total score on all of the quizzes, and

[18] [18] 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.

[19] In the data analysis below we disaggregate the CRT and Takeover game quizzes and treat them separately. The rationale for combining them in the experiment was to prevent a single question (the Takeover quiz) from having an excessively disproportionate weight. Also, in our analysis we use a score for the Takeover game that is linearly decreasing in a subject's bid. 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.

[20] As we discuss later, in Result 4, we conducted a second iteration of the experiment in which we reversed the order of the two families of games. The results are very similar, indicating that the order of games did not affect the results.

Game ID	Game Type	Player's Limits & Target	Opponent's Limits & Target
UG1	Undercutting Game	See Figure 1	
UG2	Undercutting Game	See Figure 2	
UG3	Undercutting Game	See Figure 3	
UG4	Undercutting Game	See Figure 4	
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 1. The ten games used in the experiment.

against the opponent (other than themselves) with the lowest score on the quizzes. All choices were made without feedback. After making these three choices in all ten games, players learned that they could "loop back" through the games to revise their choices if desired. This could be done up to four times, for a total of five iterations through the ten games, all without feedback.[21]

Once subjects finished all five iterations—or declined the opportunity to loop back—their play was recorded, four of their choices were randomly selected (two from the undercutting games and two from the guessing games), and they were matched with another player and paid for their decisions. Subjects earning less than $\$6$ (the standard show-up fee) were paid $\$6$ for their time. Subjects earned an average of $\$24.85$ overall.

6 DATA ANALYSIS PROCEDURES

Each subject played ten games, each against three different opponents, for a total of thirty game-play observations per subject. We employed three signals ($T = \{\tau^{\text{LO}}, \tau^0, \tau^{\text{HI}}\}$) indicating, respectively, whether the opponent had the lowest quiz score, was randomly selected, or had the highest quiz score. Following 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. We chose games so that the estimated levels (or, at least, players' relative rankings of levels) are fairly robust to these assumptions. Furthermore, the guessing-game parameters were 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.

[21] Most subjects do not use the "loop back" option. In the undercutting games 15.5% submit a final choice different from their initial choice. In the guessing games this drops to 7.2%.

For 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 player i and level k we define a likelihood function $L(s_{i\gamma\tau}|k, \lambda_i, \varepsilon_i)$ for each level k based on the assumption that player i plays the Level-k strategy $s^k_{i\gamma}$ with probability ε_i, and otherwise maximizes a random expected utility with beliefs $\nu(k)$, extreme-value-distributed noise, and sensitivity parameter λ_i. We allow both parameters (ε_i and λ_i) to differ across players.

Formally, let $\bar{s}_{i\gamma\tau}$ be the value of the observed $s_{i\gamma\tau}$ rounded to the nearest integer, and let $I_i(s_{i\gamma\tau}, k)$ be an indicator function that equals one if $\bar{s}_{i\gamma\tau} = s^k_{i\gamma}$, where s^k_i denotes the Level-k strategy for player i in game γ.[22] $I_i(s_{i\gamma\tau}, k)$ 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

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

For $k = 0$ we set $L(s_{i\gamma\tau}|0, \lambda_i, \varepsilon_i)$ equal to $\sigma^0_{i\gamma}$, 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

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

In practice, we search over a non-uniform grid of 122 possible values for λ_i and a uniform grid of 19 possible values for ε_i for each player i. The maximum-likelihood level for player i is then given by

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

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 guessing games ($G = \{5, \ldots, 10\}$). This enables us to compare stability of levels across families of games. This pooling of several

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

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

games per estimate also matches the standard procedure for estimating levels in the literature.[23] 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.[24] In the appendix, we also explore intermediate cases where two or three games per estimate are used.

7 RESULTS

Result 1: Aggregate Distributions of Levels

The distributions of levels, both for each game family and for each individual game, are shown in Table 2. The aggregate game family distributions represent fairly typical distributions of estimated levels: Level-0 is observed fairly infrequently, Level-1 is the modal type, and Level-2 and Level-3 are observed less frequently. The distribution for guessing games is similar to the distribution found by CGC06. We do find that Nash play in our undercutting games is noticeably higher than what is found in many other games.

[23] 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.

[24] 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. For the case of $|G| = 1$, we alternatively estimated levels in the guessing games by eliminating ε and $I_i(s_{i\gamma\tau}, k)$ and setting $\lambda = 1.33$ (the average estimated value of λ in CGC06 using only subjects' guesses). We then assigned a k to each observation using maximum likelihood 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.

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.[25]

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 0.86% 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.

We also find that 14.22% of all observations in the guessing games correspond exactly to one of the four (non-zero) levels' predictions, after rounding. 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, which accounts for roughly one-half of all the exact hits.[26]

In an online appendix, we compare a graphical illustration of the likelihood functions for each level with a histogram of actions. This analysis shows that there does not appear to be a substantial and regular concordance between actions and the predicted behavior of different types across games. That is, the spikes in the likelihood functions do not consistently coincide with spikes in the data for any type across the different games.[27]

[25] 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.

[26] By contrast, 48.9% of the observations in CGC06's data exactly correspond to one of the four levels' predictions. Cross-game variation in the distribution of levels remains high, however. See the online appendix for details. Again, we conjecture that these differences are due to differences in experimental instructions and their use of a best-response understanding test.

[27] This exercise also reveals purely mechanical reasons why subjects are classified more frequently as the Level-1 and Nash types. First, Level-1 beliefs are disperse, which means its likelihood function is quite flat. At the same λ_i, all higher levels have 'spike-shaped' likelihoods that exceed the Level-1 likelihood only in a small neighborhood around the predicted action. With uniformly-distributed random data, for example, Level-1 would be estimated to be the modal type for this reason. Second, the Nash type's predicted play is often at a boundary, so that logistic-response trembles can only occur in one direction. This truncation doubles the likelihood function on the interior of the strategy space, giving it a relative advantage over types with an interior prediction. We do find that estimated values of λ_i differ significantly across levels, but this appears to happen because those with noisier decisions are more likely to be classified as Level-1 due to its flatter likelihood function. In fact, the same correlation between λ_i and k is found when estimated on randomly-generated strategy data. This suggests that these two parameters do not capture independent traits—a Level-2 subject who becomes noisier is likely to be re-classified as Level-1—and that the correlation between them should not be interpreted as an insightful result.

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 3. Markov transitions from the pooled undercutting games to the pooled guessing games.

Result 2: Persistence of Absolute Levels

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 of levels between the two families of games. Table 3 reports the frequency with which a subject moves from each level in the pooled undercutting games to each level in the pooled guessing games. From the table, it is apparent that most of the transitions are into Level-1 and Nash types in the guessing game, and that these transitions do not show great correlation with a subject's type in the undercutting games. The distributions in separate rows of Table 3 are generally similar to the overall distribution in the final row, which would occur if types were independent across families of games.

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 any row. In Table 3 the overall frequencies of the levels would imply a 29.4% CLPA under the null hypothesis of independent, randomly-drawn levels. The actual CLPA is 27.3%, suggesting a slight *negative* correlation in types across games.

To test whether levels are uncorrelated, we generate 10,000 samples of 116 randomly-drawn levels, with each sample drawn independently using the overall distribution from Table 3. For each sample we calculate the CLPA, generating an approximate distribution of CLPA values under the null hypothesis. A comparison of the actual CLPA with this distribution fails to reject the null that levels are randomly drawn across game families (p-value 0.68).[28]

[28] 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 61, p.225] and comparing it against the null hypothesis of independently-drawn levels, which would give an expected Cramér correlation of zero. When

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 4. 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 5. Markov transition between single-game levels within the six guessing games.

Tables 4 and 5 also 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 the undercutting games, over half of all Level-1, Level-2 and Nash types keep the same type across games. In the undercutting games, the overall frequencies of the levels (given in the last row of Table 4) 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 perfectly accurate. In a Monte Carlo simulation of 10,000 samples of independently-drawn levels, none have a CLPA this large. Thus, we reject the null hypothesis of random levels with a p-value of less than 0.0001.

The results are quite different in the guessing games, where Level-1 acts as an absorbing state. Little difference is seen between the rows of Table 5, suggesting no correlation across games. The realized prediction accuracy (CLPA) is 34.7%. The expected CLPA under randomly-drawn levels is 31.1%. Our Monte Carlo simulation of randomly-generated levels does reject the null hypothesis with a p-value of 0.0030, though the absolute magnitude of

comparing between the two families of games using pooled-game estimates (Table 3), 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.

the difference (34.7% versus 31.1%) implies little real gain in predictive accuracy over the assumption of random levels.[29]

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 may be applied robustly in some settings, but not in others. Little guidance is currently available as to which families of games will trigger Level-k reasoning and which will not. In short, using a player's level in one game to predict her action in another may be a futile exercise without further information about the factors that determine whether Level-k reasoning is triggered.

Result 3: Persistence of Relative Levels

To examine the frequency with which the ordinal ranking of players' levels changes between the two families of games, we consider 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," or the frequency with which the same player has a strictly higher level in both games ($k_i(\gamma, \tau) > k_j(\gamma, \tau)$ and $k_i(\gamma', \tau) > k_j(\gamma', \tau)$). Pairs whose levels are the same in at least one game are excluded, so the switch and non-switch frequencies often do not sum to one. 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.[30]

The switch frequency, non-switch frequency, and switch ratio when comparing the pooled undercutting games to the pooled guessing games are reported in Table 6. The table also reports these statistics for the four undercutting games, and the six guessing games. The last column shows the predicted values under the null hypothesis of independently-drawn levels.

For the comparison between game families, switching actually occurs more frequently than non-switching. In other words, if Anne exhibits a higher level than Bob in the undercutting games, then Bob is more likely to have a higher level in the guessing games. Our 10,000-sample Monte Carlo simulation actually rejects the null hypothesis in favor of *negatively* correlated levels, with a p-value of 0.0230. This is consistent with our earlier observation that absolute levels are negatively correlated across families of games.

[29] Our analysis of the CGC06 data (in the online appendix) reveals a CLPA of 41.9%, which is between that of our guessing games and our undercutting games.

[30] In practice, the switch ratio would not exactly equal zero since some Level-0 players would be incorrectly identified as higher-level players. Our simulations suggest the actual switch ratio would be around 0.09 using our overall level distributions.

	Data	Null Hyp.
Pooled UGs vs. Pooled GGs		
Switch Frequency:	25.0%	24.9%
Non-Switch Frequency:	22.7%	24.9%
Switch Ratio:	1.10	1.00
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

TABLE 6. Observed frequency with which two players' levels strictly switch their ordering, compared to the expected frequency under independent, randomly-drawn levels.

Since absolute levels within 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. None of the 10,000 simulated samples have a switch ratio this low, indicating a clear rejection of the null hypothesis at the 0.0001 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 yields a marginal p-value of exactly 0.05. Thus, while there is some stability of relative levels within the guessing games, it is much weaker than the stability we observe in undercutting games, both in magnitude and statistical significance.

Overall, we conclude that little to 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 4: Robustness to Order of Play

In our experiment, every subject played the games in the same order, with the undercutting games (UG1–UG4) first and the guessing games (GG5–GG10) second. We find conformance with the Level-k model in the first set of games, but not in the second. One explanation for this result is that subjects become fatigued or lose attention through the course of the experiment, leading to more random play in later games. To test this hypothesis, we ran two new sessions in which the order of the games was completely reversed. Twenty-eight Ohio State subjects participated in two sessions of fourteen subjects each, using slightly modified versions of our original software and instrucitons.[31] Participants took the quizzes, played GG10 through GG5, and then played UG4 through UG1.

Recall that in our original data, the CLPA within the family of undercutting games is 57.6%, which is significantly greater than the 26.3% expected under randomly-drawn levels. In our new experiment, where the undercutting games are played second, the CLPA rises to 62.5%, indicating a slightly stronger conformance with the Level-k model. Again we reject the null hypothesis ($p < 0.0001$). The switch ratio, however, rises from 0.29 in the original data to 0.37 in the new data, indicating slightly weaker conformance with the Level-k model. But again we reject the null hypothesis of random levels with $p < 0.0001$. So the overall conclusion remains the same: across-game behavior in the undercutting games is fairly consistent with the Level-k model, and certainly far from behavior expected under the hypothesis of randomly-drawn levels, even when these games are played second.

For the guessing games, the CLPA shifts from 34.7% in our original data to 38.3% when these games appear first. This suggests slightly stronger conformance with the Level-k model. Again we reject the null ($p = 0.0070$), but note that the predictive accuracy is only 6.4 percentage points higher than under randomly-generated levels. The switch ratio, however, rises from 0.89 in the original data to 0.93 in the reversed treatment, indicating less conformance with the Level-k model. Again we cannot reject the null hypothesis ($p = 0.281$).

In summary, we do not find that overall conformance with the Level-k model is significantly improved in the guessing games when they appear first, or worsened in the undercutting games when they appear second.[32] Thus, we reject the hypothesis that order effects explain our results.

[31] To determine how many subjects were needed, we ran a bootstrapped power calculation for the Monte Carlo test of switch ratios. Specifically, for various values of $\hat{n} < 116$, we created 10,000 simulated data sets by drawing $\hat{n}$ subjects (with replacement) from our original data. Then we generated 10,000 samples of $\hat{n}$ simulated subjects with randomly-chosen levels. The power of the Monte Carlo test at $\hat{n}$ is the fraction of "real" data sets whose switch ratio is less than 95% of the "random" data sets. The usual minimum power requirement of 80% is achieved at $\hat{n} = 9$. To be conservative, we aimed to recruit at least 20 subjects, and actually had 28 participate. Our test power is over 99%.

[32] Game-by-game level distributions look similar to the original data, though the reversed data exhibit slightly more Level-1 and Level-2 play and less Level-3 and Nash play. The distributions are fairly stable in the undercutting games and highly variable across guessing games. The between-family CLPA rises to 41.0% in the reversed treatment, but we still cannot reject the null of random levels ($p = 0.112$). The between-family switch ratio drops noticeably from 1.10 to 0.76, but again we cannot reject the hypothesis of random levels ($p = 0.308$).

	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 7. Regression of expected earnings on the five quiz scores.

Result 5: Using Quizzes to Predict Levels

We next consider whether the five quizzes we administered as potential independent measures of strategic sophistication—the IQ quiz, the Eye Gaze quiz, a memory quiz, the Cognitive Reflection Test (CRT), and a one-player Takeover Game—predict behavior in the games. Because levels in the guessing games are very unstable, we do not expect them to be predictable by quiz scores. But levels in the undercutting game are stable, so we are particularly interested in whether these levels can be predicted using the quizzes.

First we examine correlations between scores on the various quizzes. These are surprisingly weak. IQ, memory, and CRT scores all appear to be positively correlated, though their estimated Spearman rank correlation coefficients achieve only marginal significance. No other correlations are statistically significant. One might conjecture that our subjects did not exert sufficient effort on the quizzes, leading to noisier scores, but absolute performance seems in line with previous studies for all quizzes except the Takeover Game.[33] Thus, we do not believe the lack of correlation is caused by unusually poor performance or lack of effort. Instead, we believe these quizzes measure relatively orthogonal traits.

Next we ask whether quiz scores predict overall earnings. To reduce randomness in the earnings measure, we calculate what each subject's expected earnings would be in each game if they played against the empirical distribution of actions of all other subjects. The correlation between subjects' total expected earnings 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 7) reveals that only the Cognitive Reflection Test (CRT) score is significantly correlated with expected earnings.

Intuitively, players using higher levels should be more sophisticated. But they do not earn more money. Indeed, Level-2 is the most profitable type, since most subjects are estimated to

[33] In the Eye Gaze test, [7] report that the average score in the general population is 81%. Our subjects scored 80% on average. In the Wechsler digit span quiz, the average number of digits correctly recalled before the first failure is 5.63. In clinical applications the test stops after *two* failures; the average score among normal adults is between 5 and 7 [21, p.416], consistent with our results. In the CRT, the percentage of players scoring $(0, 1, 2, 3)$ (respectively) is $(36\%, 28\%, 22\%, 14\%)$, which is very close to the overall average of $(33\%, 28\%, 23\%, 17\%)$ reported in Frederick's (2005) meta-study. In the Takeover Game, our subjects performed worse than in past studies: the mean bid was 94.3 for our subjects, while most studies report mean bids around 50 [39, e.g.]. We do not have comparable data for our Mensa IQ test scores.

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 8. 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.

be Level-1 types. When looking at correlations between quiz scores and estimated levels, we therefore do not restrict ourselves to a linear relationship, as Level-2 types may actually score the highest on the quizzes.

We focus on predicting levels estimated from the pooled families of games. 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 is diagnostic of 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 quiz scores. Since multinomial logistic regressions require an omitted level against which all others are compared, one single regression is not useful in analyzing all possible comparisons. We therefore report the coefficient estimates from all five possible regressions, where each regression omits a different level.[34]

Figure 6 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 8) 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.

[34] 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).

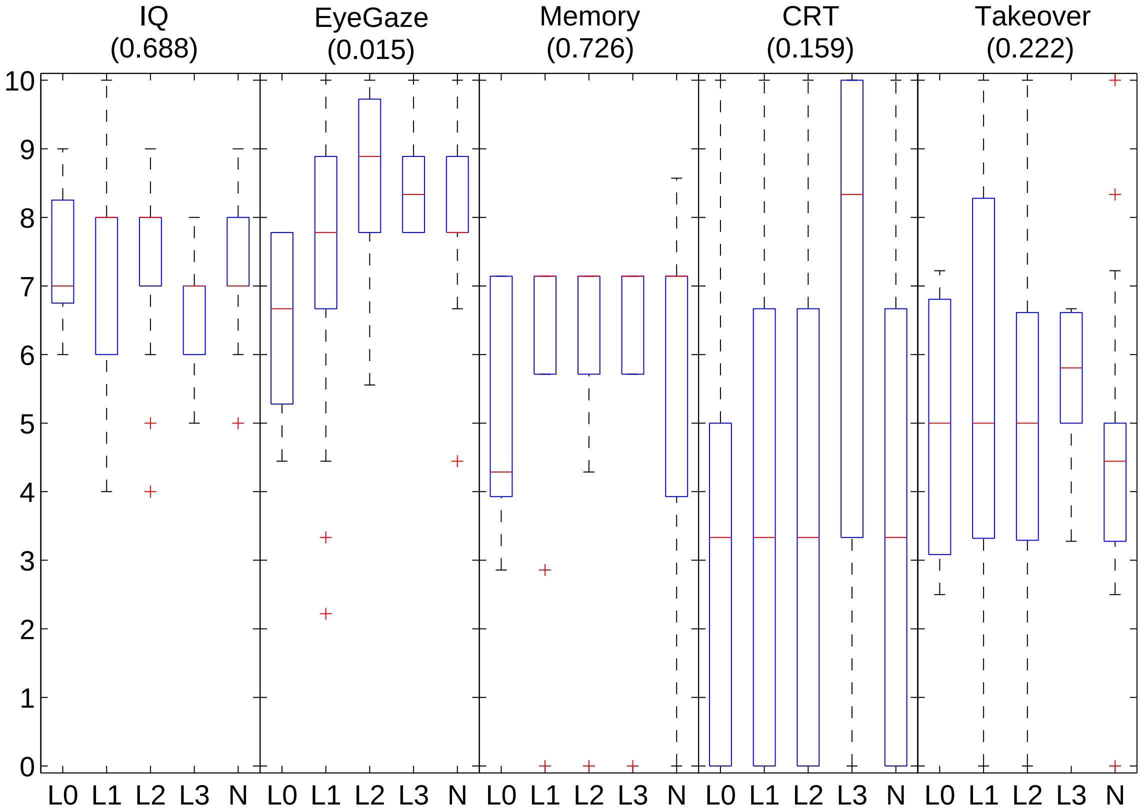


FIGURE 6. Box plots of quiz scores for each estimated level in the pooled undercutting games. p-values in parentheses are for Kruskal-Walis tests that all levels generate the same distribution of quiz scores.

The Eye Gaze correlation with Level-0 and Level-1 play has intuitive appeal: Poor performance on the Eye Gaze quiz is diagnostic of adult autism [6]. And autism is often characterized the absence of "theory of mind" [5], or an inability to recognize that others behave in response to conscious thought. This suggests that some of the Level-0 and Level-1 types are less able to consider others' beliefs and strategies in games, leading them to play more low-level actions.[35]

Figure 7 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 9) indicate that Level-2 can be distinguished from the two higher levels, but not from the two lower levels.

[35] In p-beauty contest games, [22] find that higher-level players exhibit greater neural activation in the medial prefrontal cortex (mPFC). Theory-of-mind experiments also find activation in this region (among others). These results are roughly consistent with our Eye Gaze finding, and also suggest more predictable heterogeneity in p-beauty contest games.

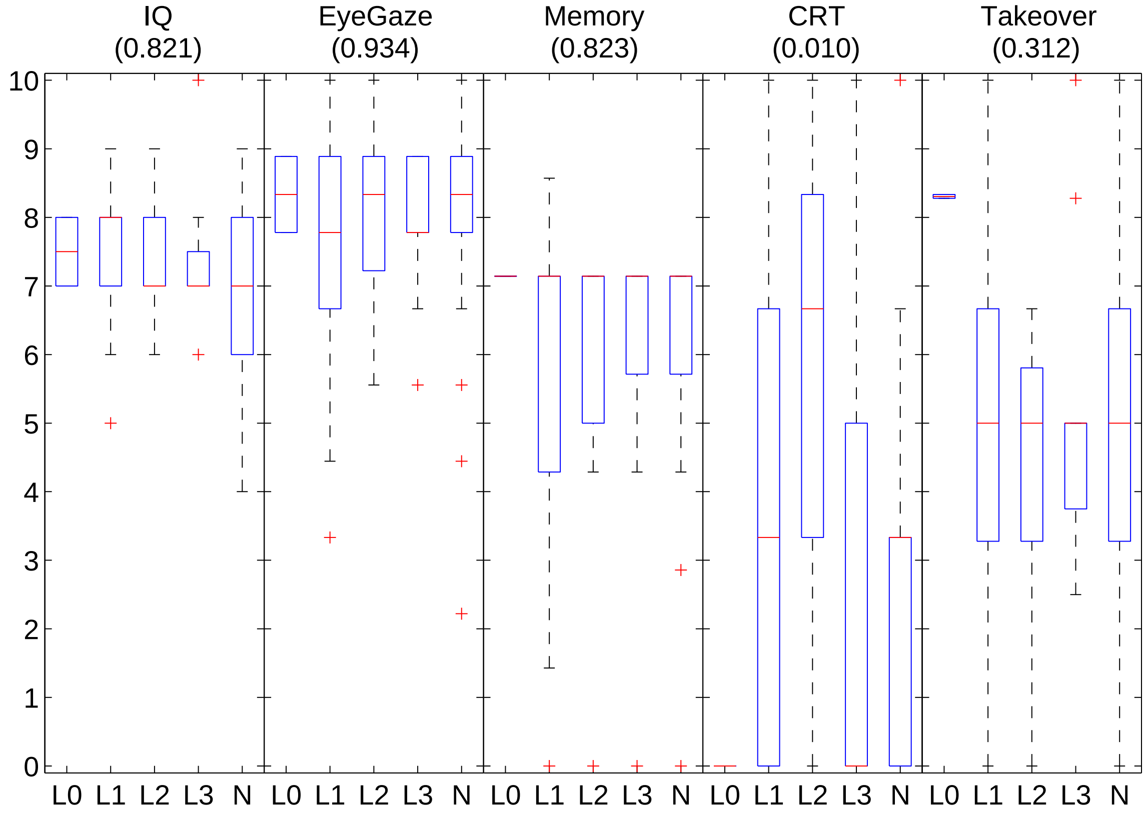


FIGURE 7. Box plots of quiz scores for each estimated level in the pooled guessing games. p-values in parentheses are for Kruskal-Walis tests that all levels generate the same distribution of quiz scores.

If levels in the guessing games are unstable, how can the CRT quiz be predictive of Level-2? Because Level-2 is also a proxy for higher earnings. We already know that the CRT quiz predicts earnings (Table 7), and those who earn more are more likely to be classified as Level-2, so the correlation between CRT quizzes and Level-2 appears spurious.

We perform similar analyses for game-by-game levels, and the results are consistent with the pooled-game results. In the undercutting 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 is greater than 0.15 for all other quizzes.

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

TABLE 9. Multinomial logit regression coefficient estimates of CRT quiz scores on pooled guessing game levels. Each column represents a regression with a different omitted category.

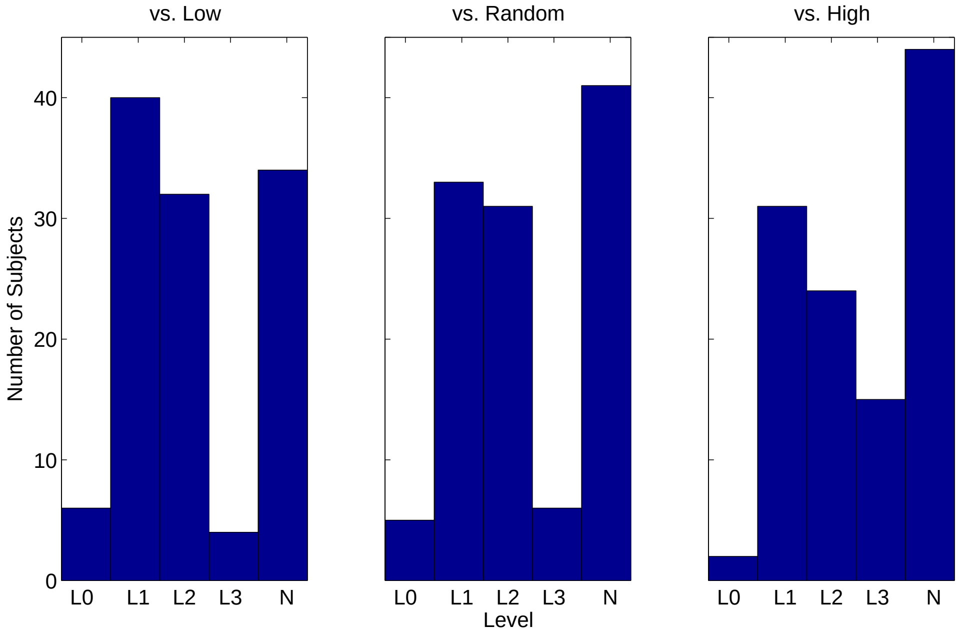


FIGURE 8. Level distributions by opponent in the pooled undercutting games.

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 opponent with the highest total quiz score, and against the opponent 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

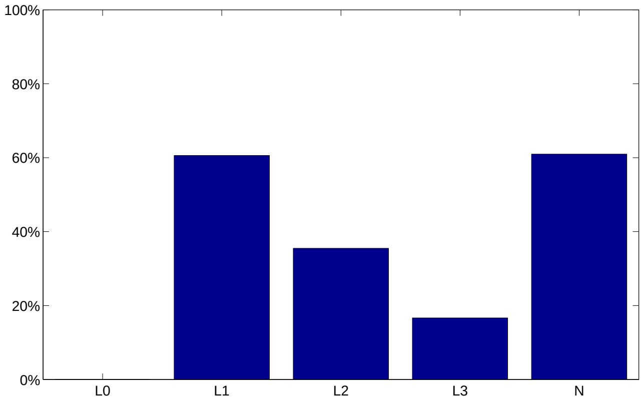


FIGURE 9. For each k, the percentage of Level-k subjects in pooled undercutting games who do not change levels in response to opponent types.

that subjects might treat quiz scores as proxies for strategic sophistication.[36] Thus, how subjects respond to their opponents' characteristics may provide another testable prediction for the Level-k model.

Figure 8 shows the histogram of estimated levels in the pooled undercutting games for each of the three types of opponent. 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. χ^2 tests confirm that the distribution of levels is significantly different between the low-score and high-score opponent (p-value of 0.018), though not significantly different between the low-score and random opponents (p-value of 0.767) or between the random and high-score opponents (p-value of 0.185).

While the above differences in behavior by opponent are interesting, we are concerned with whether any information can predict this adjustment. That is, can we predict which subjects have a high enough "capacity" to be able to adjust their behavior in response to information about opponents? We therefore ask whether quiz scores 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 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.

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

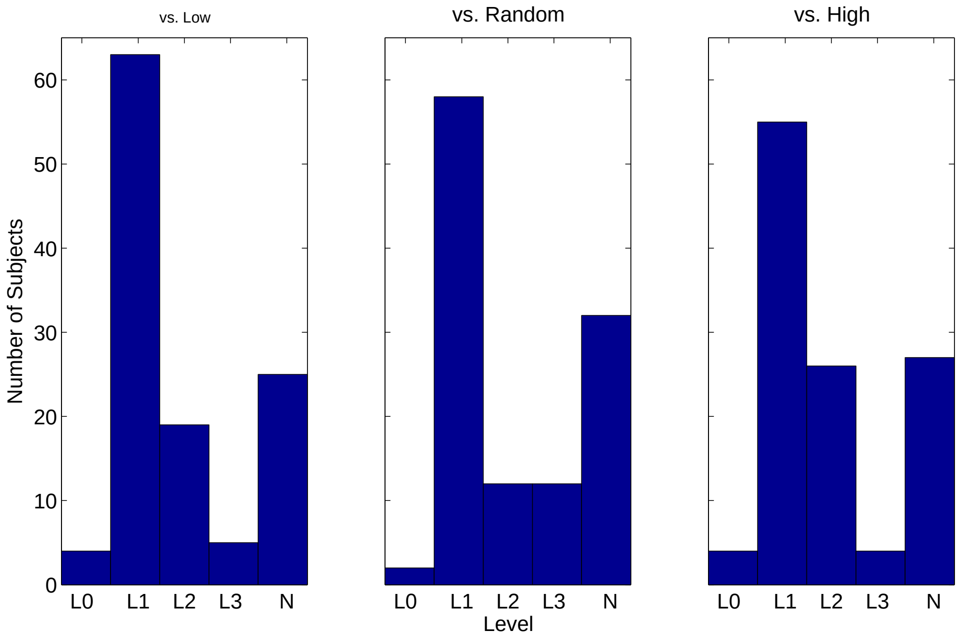


FIGURE 10. Level distributions by opponent in the pooled guessing games.

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 against random opponents whose levels do not shift in response to high- or low-scoring opponents. We refer to these as *stable* players. The percentage of stable players for each k are shown in Figure 9. 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. Thus, there may exist heterogeneity amongst players beyond the number of best responses they perform.

Similar analyses in the pooled guessing games (Figure 10) yields no significance differences in the low-vs-random and low-vs-high comparisons (χ^2 p-values of 0.185 and 0.769, respectively). We do find a significant difference in level distributions between random opponents and high-scoring opponents (p-value of 0.035), but the mean level against high-scoring opponents

	Frequency	i.i.d. Prob.
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:	1.078	1.00
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

TABLE 10. 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.

is actually lower than against random opponents (1.957 versus 2.121), and a Wilcoxon-Mann-Whitney test reveals no stochastic dominance of these level distributions (p-value 0.541), so we cannot claim that players' levels unambiguously shift up (or down) against stronger opponents.[37]

In summary, we do see some subjects adjusting their realized levels against different opponents, particularly in the undercutting games. This indicates 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 10 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 are compared against

[37] [49] shows that this test can be viewed as a test of stochastic dominance, even with discrete distributions. [33] provide an excellent survey of valid perspectives for this test.

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

Comparing the pooled undercutting games with the pooled guessing games, switches occur more often in the opposite direction than in the same direction. The ratio of switch directions is close to 1, which is what one would expect if levels were independently drawn in each game family. A Monte Carlo simulation with 1,000 samples shows that the empirical ratio of switch directions fails to reject the null hypothesis of independently-drawn levels, with a p-value of 0.380. Thus, the two families of games cannot be clearly ranked using estimated levels.

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 shows 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—the Monte Carlo simulation yields a p-value of 0.070. This occurs because the level distributions vary more across guessing games, so the variance of switch directions under the null hypothesis is much larger.

8 Discussion

As a broad summary of our findings, the success of the Level-k model is mixed: We find very little cross-game stability when comparing the family of undercutting games with the family of two-person guessing games. We do find reasonably strong cross-game stability within the family of undercutting games, but zero stability in the two-person guessing games. Even in the undercutting games, however, observed levels are hard to predict with our five psychometric measures, except that Level-1 players may have a less keen awareness of others' emotions and cognition. Finally, it appears that some players "step up" against stronger opponents in undercutting games, 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 [62, 63, 54, 31, 42, 10, 18] or behavioral data augmented with lookup data [24, 23] or eye-tracking data [19, 65]. For some games, however, the Level-k model does not appear to organize the data well [43, 44, 26].[38] [60] 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 [52]. The broad conclusion that emerges from this line of research is that the Level-k approach works well in some games, but not in others.

Camerer et al. [18, 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. This suggests that the Level-k model may be one of many possible decision processes players employ to select strategies in novel games. Different processes may be triggered unconsciously in different settings, depending on features such as the characteristics of the game and the way in which the game is described. 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, Level-k reasoning may not be triggered unless subjects are given sufficient instruction and experience calculating best responses prior to play (see the online appendix).

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 other settings we expect less frequency of Level-k reasoning. 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 distinct decision "heuristics" employed by people playing games *and* exploring their triggers. For example, [43] identify plausible "rules of thumb" to explain their data when Level-k and quantal

[38] [26] point out that the Level-k model fails to account for overbidding in second-price auctions.

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.

Finally, a 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.

References

[1] Steffen Altmann and Armin Falk. The impact of cooperation defaults on voluntary contributions to public goods. Univeristy of Bonn Working Paper, 2009.

[2] Ayala Arad and Ariel Rubinstein. The 11-20 money request game: A level-k reasoning study. *American Economic Review*, 102(7):3561–73, 2012.

[3] Ayala Arad and Ariel Rubinstein. Multi-dimensional iterative reasoning in action: The case of the Colonel Blotto game. *Journal of Economic Behavior and Organization*, 84: 571–585, 2012.

[4] S. B. Ball, Max H. Bazerman, and J. S. Carroll. An evaluation of learning in the bilateral winner's curse. *Organizational Behavior and Human Decision Processes*, 48:1–22, 1991.

[5] Simon Baron-Cohen. Autism: A specific cognitive disorder of 'mind-blindness. *International Review of Psychiatry*, 2(1):81–90, 1990.

[6] Simon Baron-Cohen, Therese Jolliffe, Catherine Mortimore, and Mary Robertson. 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, 1997.

[7] Simon Baron-Cohen, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb. The "reading the mind in the eyes" test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, 42:241–251, 2001.

[8] D. Batzilis, S. Jaffe, J.A. Levitt, J List, and J. Picel. Large-scale analysis of level-k thinking: Facebooks Roshambull. University of Chicago Working Paper, 2013.

[9] Meghana Bhatt and Colin F. Camerer. Self-referential thinking and equilibrium as states of mind in games: fMRI evidence. *Games and Economic Behavior*, 52(2):424–459, 2005.

[10] Antoni Bosch-Domènech, Jose Garcia-Montalvo, Rosemarie C. Nagel, and Albert Satorra. One, two, (three), infinity...: Newspaper and lab beauty-contest experiments. *American Economic Review*, 92(5):1687–1701, 2002.

[11] Isabelle Brocas, Colin Camerer, Juan D. Carrillo, and Stephanie W. Wang. Measuring attention and strategic behavior in games with private information. CEPR Discussion Paper No. DP7529, 2009.

[12] Alexander L. Brown, Colin F. Camerer, and Dan Lovallo. To review or not to review? Limited strategic thinking at the movie box office. *American Economic Journal: Microeconomics*, 4:1–26, 2012.

[13] Antoine J. Bruguier, Steven R. Quartz, and Peter L. Bossaerts. Exploring the nature of "trading intuition". *Journal of Finance*, 65:1703–1723, 2010. California Institute of Technology working paper.

[14] K. Burchardi and S. Penczynski. Out of your mind: Eliciting individual reasoning in one shot games. Games and Economic Behavior, forthcoming., 2011.

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

[16] Colin F. Camerer. *Behavioral Game Theory*. Princeton University Press, Princeton, NJ, 2003.

[17] Colin F. Camerer and Teck-Hua Ho. Experience-weighted attraction learning in coordination games: Probability rules, heterogeneity and time-variation. *Journal of Mathematical Psychology*, 42:305–326, 1998.

[18] Colin F. Camerer, Tech-Hua Ho, and Juin-Kuan Chong. A cognitive heirarchy model of games. *Quarterly Journal of Economics*, 119(3):861–898, 2004.

[19] Chun-Ting Chen, Chen-Ying Huang, and Joseph Tao-yi Wang. A window of cognition: Eyetracking the reasoning process in spatial beauty contest games. National Taiwan University working paper, 2009.

[20] Juin-Kuan Chong, Colin F. Camerer, and Teck-Hua Ho. Cognitive hierarchy: A limited thinking theory in games. In Rami Zwick and Amnon Rapoport, editors, *Experimental Business Research: Marketing, Accounting and Cognitive Perspectives*, volume III, chapter 9, pages 203–228. Springer, The Netherlands, 2005.

[21] Ronald A. Cohen, Stephen Salloway, and Lawrence H. Sweet. Neuropsychiatric aspects of disorder of attention. In *American Psychiatric Press Textbook of Neuropsychiatry*, chapter 10, pages 405–444. American Psychiatric Publishing, Arlington, Virginia, 2007.

[22] Giorgio Coricelli and Rose Nagel. Neural correlates of depth of strategic reasoning in medial prefontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, DOI: 10.1073/pnas.0807721106, 2009.

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

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

[25] Vincent P. Crawford. Adaptive dynamics in coordination games. *Econometrica*, 63: 103–143, 1995.

[26] Vincent P. Crawford and Nagore Iriberri. 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, 2007.

[27] Vincent P. Crawford and Nagore Iriberri. Fatal attraction: Salience, naivete, and sophistication in experimental "hide-and-seek" games. *American Economic Review*, 97(5): 1731–1750, 2007.

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

[29] J. De Sousa, G. Hollard, and A. Terracol. Cognitive ability and learning to play equilibrium: A level-k analysis. Working paper., 2012.

[30] Giovanna Devetag and Massimo Warglien. Games and phone numbers: Do short-term memory bounds affect strategic behavior? *Journal of Economic Psychology*, 24:189–202, 2003.

[31] John Duffy and Rosemarie C. Nagel. On the robustness of behavior in experimental 'beauty contest' games. *Economic Journal*, 107:1684–1700, 1997.

[32] Ido Erev and Alvin E. Roth. Prediction how people play games: Reinforcement learning in games with unique strategy equilibrium. *American Economic Review*, 88:848–881, 1998.

[33] Michael P. Fay and Michael A. Proschan. Wilcoxon-mann-whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, 4:1–39, 2010.

[34] Shane Frederick. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19:25–42, 2005.

[35] Sotiris Georganas. English auctions with resale: An experimental study. *Games and Economic Behavior*, 73:147–166, 2011.

[36] Gerd Gigerenzer. The adaptive toolbox. In Gerd Gigerenzer and Reinhard Selten, editors, *Bounded Rationality: The Adaptive Toolbox*, chapter 3. MIT Press, London, 2001.

[37] David Gill and Victoria Prowse. Cognitive ability and learning to play equilibrium: A level-k analysis. Oxford University Working Paper, 2013.

[38] Brit Grosskopf and Rosemarie Nagel. The two-person beauty contest. *Games and Economic Behavior*, 62:93–99, 2008.

[39] Grosskopf Grosskopf, Yoella Bereby-Meyer, and Max H. Bazerman. On the robustness of the winner's curse phenomenon. *Theory and Decision*, 63:389–418, 2007.

[40] John C. Harsanyi. Games with incomplete information played by "Bayesian" players, I-III. Part I: The basic model. *Management Science*, 14:159–182, 1967.

[41] Teck-Hua Ho and Xuanming Su. A dynamic level-k model in sequential games. *Management Science*, 59:452–469, 2013.

[42] Teck-Hua Ho, Colin F. Camerer, and Keith Weigelt. Iterated dominance and iterated best response in p-beauty contests. *American Economic Review*, 88:947–969, 1998.

[43] Asen Ivanov, Dan Levin, and James Peck. 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, 2009.

[44] Asen Ivanov, Dan Levin, and Muriel Niederle. Can relaxation of beliefs rationalize the winner's curse?: An experimental study. *Econometrica*, 78:1435–1452, 2010.

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

[46] Toshiji Kawagoe and Hirokazu Takizawa. Level-k analysis of experimental centipede games. *Journal of Economic Behavior and Organization*, 82:548–566, 2012.

[47] John M. Keynes. *The General Theory of Interest, Employment and Money*. Macmillan, London, 1936.

[48] Terri Kneeland. Coordination under limited depth of reasoning. University of British Columbia Working Paper, July 2012.

[49] Erich Leo Lehmann. Consistency and unbiasedness of certain nonparametric tests. *Annals of Mathematical Statistics*, 22:165–179, 1951.

[50] Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal-form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

[51] Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for extensive-form games. *Experimental Economics*, 1:9–41, 1998.

[52] Stephen Morris and Hyun Song Shin. Social value of public information. *American Economic Review*, 92:1521–1534, 2002.

[53] Rosemarie C. Nagel. Experimental results on interactive competitive guessing. Discussion Paper 8-236, Sonderforschungsbereich 303, Universitat Bonn, 1993.

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

[55] Robert Ostling, Joseph Tao-yi Wang, Eileen Chou, and Colin F. Camerer. Testing game theory in the field: Swedish lupi lottery games. *American Economic Journal: Microeconomics*, 3:1–33, 2011.

[56] Stefan Penczynski. Strategic thinking: The influence of the game. Working paper, 2011.

[57] Brian W. Rogers, Thomas R. Palfrey, and Colin F. Camerer. Heterogeneous quantal response equilibrium and cognitive heirarchies. *Journal of Economic Theory*, 144:1440–1467, 2009.

[58] Ariel Rubinstein. A typology of players: Between instinctive and contemplative. Tel Aviv University Working Paper, 2014.

[59] William F. Samuelson and Max H. Bazerman. The winner's curse in bilateral negotiations. In Vernon L. Smith, editor, *Research in Experimental Economics*, volume 3, pages 105–137. JAI Press, Greenwich, CT, 1985.

[60] Dmitry Shapiro, Xianwen Shi, and Artie Zillante. Robustness of level-k resaoning in generalized beauty contest games. University of North Carolina working paper, October 2009.

[61] Sidney Siegel and N. John Castellan, Jr. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, NY, 2nd edition, 1988.

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

[63] Dale O. Stahl and Paul W. Wilson. On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10:218–254, 1995.

[64] Tomasz Strzalecki. Depth of reasoning and higher-order beliefs. Harvard University Working Paper, 2009.

[65] Joseph Tao-yi Wang, Michael Spezio, and Colin F. Camerer. 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, 2009.

[66] David Wechsler. *Measurement of adult intelligence*. Williams & Wilkins, Baltimore, 1939.

Acknowledgments:

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. Financial support for this research was provided by the Ohio State University Department of Economics. The authors have no relevant or material financial interests that relate to the research described in this paper. Institutional Review Board (IRB) approval was obtained for this study.

Appendices For Online Publication

Appendix A (Online): Comparison With CGC06

The two-person guessing games used in our experiment were taken from [23]. In this appendix we compare our results to data from their procedurally similar 'Baseline' and 'Open Boxes' treatments 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).[39]

To our knowledge, there are two notable differences between our experimental design and theirs.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. Both subject pools appear to be standard within the experimental economics literature.

Second, and perhaps more 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 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.[40] 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.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.

Table 11 shows that the aggregate distribution of levels among CGC06's guessing games (estimated game-by-game) looks similar to that found in our data, though with more Level-2 subjects. But, as in our data, the game-by-game frequencies of levels feature a large degree

[39] 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.

[40] 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 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.

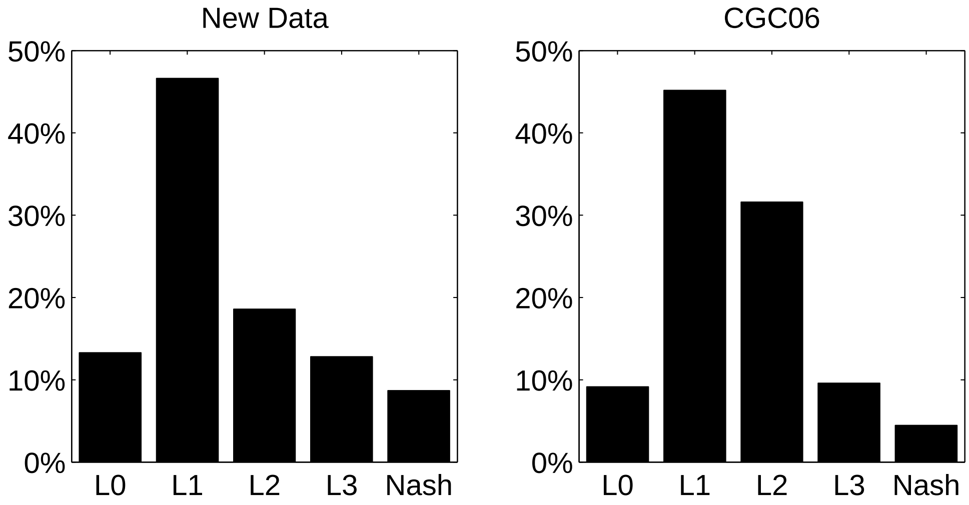


FIGURE 11. Aggregate level distributions across all standard guessing games in our data and in CGC06.

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

of heterogeneity across games. 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). Following CGC06, these 16 games are ordered so that lower-numbered games require fewer rounds of dominance elimination to solve the

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 12. Frequency of exact conformity with the Level-k model in the new data and in [23].

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 13. Markov switching matrix of levels in the CGC06 data.

equilibrium. None of the five levels' 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 12). 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 their use of a best-response understanding test, either of which may trigger a Level-k heuristic in subjects.

The stability of levels appears slightly higher in the CGC06 data, but not as stable as we found in our guessing games. The Markov transition matrix between games is shown in Table 13. 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%. 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.

	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 14. 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.

The stability of relative levels in CGC06's data also lies between that of our guessing games and our undercutting games. Table 14 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.

The improvement in stability in the CGC06 data is likely due to the lengthier instructions and the use of an understanding test. [28] 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 amount of experience 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.

APPENDIX B (ONLINE): 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 12. 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 is visible in Figure 12. 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 2. 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

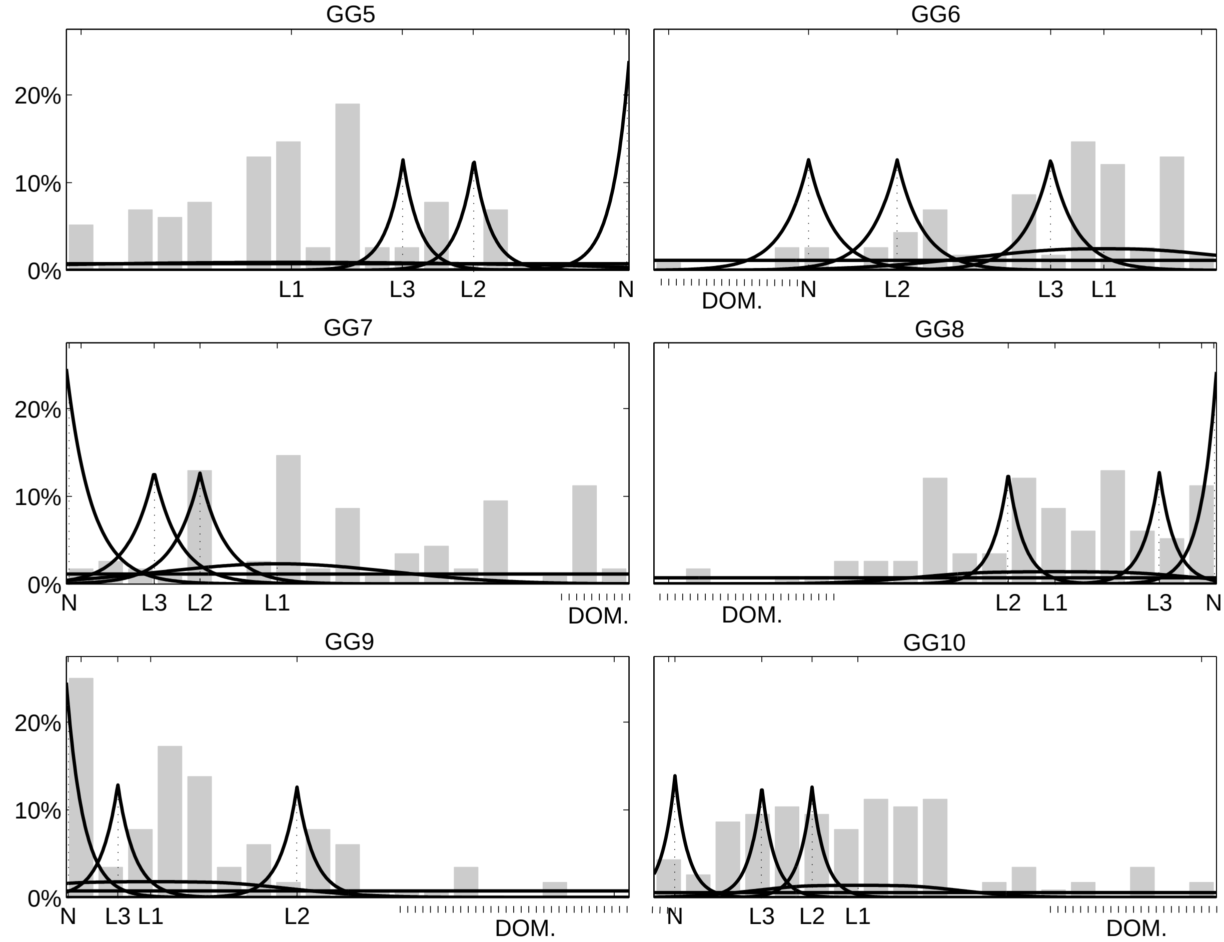


FIGURE 12. Histograms of actions in each guessing game, with likelihood functions for each level (assuming $\lambda = 1$).

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

APPENDIX C (ONLINE): 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.[41] 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

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

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 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 $m = 16$).

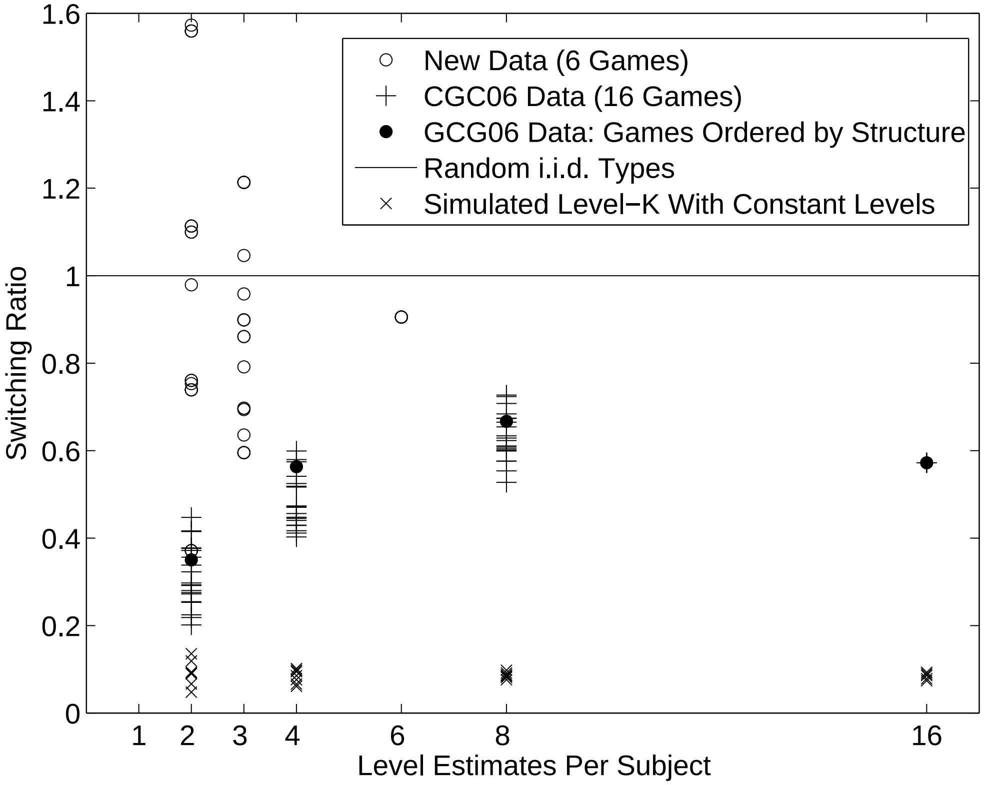


FIGURE 13. Switching ratios as the number of levels estimated per subject varies, using many randomly-drawn partitions of the games.

The results of this analysis appear in Figure 13. 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.

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.