Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)

Beyond Equilibrium: Predicting Human Behavior in Normal-Form Games

James R. Wright and **Kevin Leyton-Brown**
Department of Computer Science, University of British Columbia
2366 Main Mall, Vancouver, B.C., Canada, V6T 1Z4
{jrwright,kevinlb}@cs.ubc.ca

Abstract

It is standard in multiagent settings to assume that agents will adopt Nash equilibrium strategies. However, studies in experimental economics demonstrate that Nash equilibrium is a poor description of human players' initial behavior in normal-form games. In this paper, we consider a wide range of widely-studied models from behavioral game theory. For what we believe is the first time, we evaluate each of these models in a meta-analysis, taking as our data set large-scale and publicly-available experimental data from the literature. We then propose modifications to the best-performing model that we believe make it more suitable for practical prediction of initial play by humans in normal-form games.

Introduction

This paper investigates methods for predicting human play of normal-form games. Three notes are in order. First, by "normal-form games" we mean unrepeated interactions; some literature calls this setting "initial play." Second, by "prediction" we mean accurately forecasting action choices on unseen games; thus, we are not interested in models that simply explain observed behavior. Finally, by "human play" we refer to actual play of games by motivated human subjects in experiments; however, we are interested in data sets in which play is anonymous, and hence it is impossible to model each individual player.

Perhaps the most standard game-theoretic assumption is that all participants will adopt Nash equilibrium strategies—that they will jointly behave in a way that ensures that each participant optimally responds to the others. This solution concept has many appealing properties; e.g., in any other strategy profile, one or more participants will regret their strategy choices. However, there are three key reasons why a human player might choose not to adopt such a strategy. First, she may face a computational limitation that prevents her from computing a Nash equilibrium strategy, even if the game has only one. Second, even if she can compute an equilibrium, she may doubt that her opponents can or will do so. Third, when there are multiple equilibria, it is not clear which she should expect the other participants to adopt and hence whether she should play towards one herself—even if all players are perfectly rational.

The next-most standard approach is to devise new solution concepts that overcome problems with Nash equilibrium, e.g., competitive safety strategies (Tennenholtz 2002), minimax regret equilibrium (Hyafil and Boutilier 2004), generalized strategic eligibility (Conitzer and Sandholm 2005), CURB sets (Benisch, Davis, and Sandholm 2006), and iterated regret minimization (Halpern and Pass 2009). Still other work aims to identify strategies that work well without detailed modeling of the opponent. This line of work is perhaps exemplified by the very influential series of Trading Agent Competitions (Wellman, Greenwald, and Stone 2007).

We are most interested in approaches that make explicit predictions about which actions a player will adopt, and that are grounded in human behavior. The relatively new field of *behavioral game theory* extends game-theoretic models to account for human behavior by taking account of human cognitive biases and limitations (Camerer 2003). Experimental evidence is a cornerstone of behavioral game theory, and researchers have developed many models of how humans behave in strategic situations based on experimental data.

Among these models, four key paradigms have emerged: level-k (Costa-Gomes, Crawford, and Broseta 2001) and quantal level-k (Stahl and Wilson 1994) models, the closely-related cognitive hierarchy model (Camerer, Ho, and Chong 2004), and quantal response equilibrium (McKelvey and Palfrey 1995). Although different studies consider different specific variations, the overwhelming majority of behavioral models of initial play of normal-form games fall broadly into this categorization.

One line of work from the AI literature also meets our criteria of predicting action choices and modeling human behavior (Altman, Bercovici-Boden, and Tennenholtz 2006). This approach learns association rules between agents' actions in different games to predict how an agent will play based on its actions in earlier games. We do not consider this approach in our study, as it requires data that identifies agents across games, and cannot make predictions for games that are not in the training dataset. Nevertheless, more broadly such machine-learning-based methods could be extended to our setting; investigating their performance would be an interesting line of future work.

Given the variety of behavioral models available, we can refine our focus by asking: which of these models is best for predicting human behavior in normal-form games? We

Paper	Nash	QLk	Lk	CH	QRE
(Stahl and Wilson 1994)	t	t			
(McKelvey and Palfrey 1995)	f				f
(Stahl and Wilson 1995)	f	t			
(Costa-Gomes, Crawford, and Broseta 1998)	f		f		
(Haruvy, Stahl, and Wilson 1999)		t			
(Costa-Gomes, Crawford, and Broseta 2001)	f		f		
(Haruvy, Stahl, and Wilson 2001)		t			
(Morgan and Sefton 2002)	f				p
(Weizsäcker 2003)	t				t
(Camerer, Ho, and Chong 2004)	f			p	
(Costa-Gomes and Crawford 2006)	f		f		
(Stahl and Haruvy 2008)		t			
(Rey-Biel 2009)	t		t		
(Georganas, Healy, and Weber 2010)	f		f		
(Hahn, Lum, and Mela 2010)				p	
(Camerer, Ho, and Chong 2001)				f	f
(Chong, Camerer, and Ho 2005)	f			p	p
(Crawford and Iriberri 2007)	p		p		p
(Costa-Gomes, Crawford, and Iriberri 2009)	f		f	f	f
(Rogers, Palfrey, and Camerer 2009)	f			f	f

Table 1: Existing work. A ‘p’ indicates that the study evaluated out-of-sample prediction performance for that model; a ‘t’ indicates statistical tests of training sample performance; an ‘f’ indicates comparison of training sample fit only.

conducted an exhaustive literature survey to determine the extent to which this question had already been answered. Specifically, we checked all (1698) citations to the four papers cited above using Google Scholar. We discarded superficial references, papers that simply applied one of the models to an application domain, and papers that studied repeated games. This left us with a total of 20 papers (including the four with which we began), listed in Table 1.

Overall, we found no paper that compared the predictive performance of all four models. Indeed, there were two senses in which the literature fell short of addressing this question. First, the behavioral economics literature is concerned more with *explaining* behavior than with *predicting* it. Thus, comparisons of out-of-sample prediction performance were rare. Here we describe the only exceptions: (Morgan and Sefton 2002) and (Hahn, Lum, and Mela 2010) evaluated prediction performance using held-out test data; (Camerer, Ho, and Chong 2004) and (Chong, Camerer, and Ho 2005) computed likelihoods on each individual game in their datasets after using models fit to the $n-1$ remaining games; (Crawford and Iriberri 2007) compared the performance of two models by training each model on each game in their dataset individually, and then evaluating the performance of each of these n trained models on each of the $n-1$ other individual games. Second, most of the papers compared only one of the four models (often with variations) to Nash equilibrium. Indeed, only five of the 20 studies (see the bottom portion of Table 1) compared more than one of the four key models. Only two of these studies explicitly compared the prediction performance of more than one of the four models; the remaining three performed comparisons in terms of training set fit.

In the next section we describe canonical forms of the four key models that we evaluated, and the human experimental data upon which we based our study. We then present our paper's two key contributions. First, we evaluate the quality of the behavioral predictions made by these four models. Second, we perform deeper analyses of the models, addressing four questions that arose out of our initial evaluation. Overall, we conclude that the quantal level-k model predicts initial play by humans in normal-form games significantly better than the other behavioral models. We also propose a conceptually simpler and more parsimonious model with performance roughly equivalent to quantal level-k.

Existing Behavioral Models

We begin by formally defining the four behavioral models that we studied.

Quantal Response Equilibrium

One prominent behavioral theory asserts that agents become more likely to make errors as those errors become less costly. We refer to this property as *cost-proportional errors*. This can be modeled by assuming that agents best respond *quantally*, rather than via strict maximization.

Definition 1 (Quantal best response)**.** A *(logit) quantal best-response* $QBR_i(s_{-i} \mid \lambda)$ by agent i to a strategy profile s_{-i} is a mixed strategy s_i such that

$$s_i(a_i) = \frac{\exp[\lambda{\cdot}u_i(a_i, s_{-i})]}{\sum_{a'_i} \exp[\lambda{\cdot}u_i(a'_i, s_{-i})]}, \tag{1}$$

where λ (the *precision* parameter) indicates how sensitive agents are to utility differences. Note that unlike regular best response, which is a set-valued function, quantal best response returns a single mixed strategy. ■

This gives rise to a generalization of Nash equilibrium known as the *quantal response equilibrium* (“QRE”) (McKelvey and Palfrey 1995).

Definition 2 (QRE)**.** A *quantal response equilibrium* with precision λ is a mixed strategy profile s^* in which every agent's strategy is a quantal best response to the strategies of the other agents. That is, $s^*_i = QBR_i(s^*_{-i} \mid \lambda) \forall$ agents i. ■

A QRE is guaranteed to exist for any normal-form game and non-negative precision (McKelvey and Palfrey 1995).

One criticism of this solution concept is that, although (1) is translation-invariant, it is not scale invariant. That is, while adding some constant value to the payoffs of a game will not change its QRE, multiplying payoffs by a positive constant will. This is problematic because utility functions do not themselves have unique scales (Von Neumann and Morgenstern 1944).

Level-k

Another key idea from behavioral game theory is that humans can perform only a bounded number of *iterations of strategic reasoning*. The level-k model (Costa-Gomes, Crawford, and Broseta 2001) captures this idea by associating each agent i with a level $k_i \in \{0, 1, 2, \ldots\}$, corresponding to the number of iterations of reasoning the agent is able to perform. A *level-*0 *agent* plays randomly, choosing uniformly at random from his possible actions. A *level-k agent*, for $k \geq 1$, best responds to the strategy played by level-$(k-1)$ agents. If a level-k agent has more than one best response, he mixes uniformly over them.

Here we consider a particular level-k model, dubbed Lk, which assumes that all agents belong to levels 0,1, and 2.[1] Each agent with level $k > 0$ has an associated probability ϵ_k of making an "error", i.e., of playing an action that is not a best response to the level-$(k-1)$ strategy. However, the agents do not account for these errors when forming their beliefs about how lower-level agents will act.

Definition 3 (Lk model)**.** Let A_i denote player i's action set, and $BR_i(s_{-i})$ denote the set of i's best responses to the strategy profile s_{-i}. Let $IBR_{i,k}$ denote the *iterative best response set* for a level-k agent i, with $IBR_{i,0} = A_i$ and $IBR_{i,k} = BR_i(IBR_{-i,k-1})$. Then the distribution $\pi^{Lk}_{i,k} \in \Pi(A_i)$ that the Lk model predicts for a level-k agent playing as agent i is defined as follows:

$$\pi^{Lk}_{i,0}(a_i) = |A_i|^{-1},$$

$$\pi^{Lk}_{i,k}(a_i) = \begin{cases} (1-\epsilon_k)/|IBR_{i,k}| & \text{if } a_i \in IBR_{i,k}, \\ \epsilon_k/(|A_i| - |IBR_{i,k}|) & \text{otherwise.} \end{cases}$$

In total, this model has 4 parameters: $\{\alpha_1, \alpha_2\}$, the relative proportions of level-1 and level-2 agents, and $\{\epsilon_1, \epsilon_2\}$, the error probabilities of each non-zero type. ■

Cognitive Hierarchy

The cognitive hierarchy model (Camerer, Ho, and Chong 2004), like level-k, aims to model agents with heterogeneous bounds on iterated reasoning. It differs from the level-k model in two ways. First, agent types do not have associated error rates; each agent best responds perfectly to its beliefs. Second, agents best respond to the full distribution of lower-level types, rather than only to the strategy one level below. More formally, every agent again has an associated level $m \in \{0, 1, 2, \ldots\}$. Let F be the cumulative distribution of the levels in the population. Level-0 agents play (typically uniformly) at random. Level-m agents ($m \geq 1$) best respond to the strategies that would be played in a population described by the cumulative distribution $F(j \mid j < m)$.

(Camerer, Ho, and Chong 2004) advocate a single-parameter restriction of the cognitive hierarchy model called *Poisson-CH*, in which the levels of agents in the population F are distributed according to a Poisson distribution.

Definition 4 (Poisson-CH model)**.** Let $\pi^{PCH}_{i,m} \in \Pi(A_i)$ be the distribution over actions predicted for an agent i with level m by the Poisson-CH model. Let $F \sim \text{Poisson}(\tau)$. Let $TBR_{i,m}$ be the *truncated best response set* for a level-m agent i, with $TBR_{i,m} = BR_i\left(\sum_{\ell=0}^{m-1} F(\ell)\pi^{PCH}_{-i,\ell}\right)$. Then π^{PCH} is defined as follows:

$$\pi^{PCH}_{i,0}(a_i) = |A_i|^{-1},$$

$$\pi^{PCH}_{i,m}(a_i) = \begin{cases} |TBR_{i,m}|^{-1} & \text{if } a_i \in TBR_{i,m}, \\ 0 & \text{otherwise.} \end{cases}$$ ■

(Rogers, Palfrey, and Camerer 2009) noted that cognitive hierarchy predictions often exhibit cost-proportional errors (which they call the "negative frequency-payoff deviation relationship"), even though the cognitive hierarchy model does not explicitly model this effect. This leaves open the question whether cognitive hierarchy (and level-k) predict well only to the extent that their predictions happen to exhibit cost-proportional errors, or whether bounded iterated reasoning captures an independent phenomenon.

Quantal Level-k

(Stahl and Wilson 1994) propose a rich model of strategic reasoning that combines elements of the QRE and level-k models; we refer to it as the *quantal level-k model* (QLk). In QLk, agents have one of three levels, as in Lk. Each agent responds to its beliefs quantally, as in QRE. Like Lk, agents believe that the rest of the population has the next-lower type.

The main difference between QLk and Lk is in the error structure. In Lk, higher-level agents believe that all lower-level agents best respond perfectly, although in fact every agent has some probability of making an error. In contrast, in QLk, agents are aware of the quantal nature of the lower-level agents' responses, and have a (possibly-incorrect) belief about the lower-level agents' precision.

Definition 5 (QLk model)**.** The distribution $\pi^{QLk}_{i,k} \in \Pi(A_i)$ over actions that QLk predicts for a level-k agent playing as agent i is defined as follows.

$$\pi^{QLk}_{i,0}(a_i) = |A_i|^{-1},$$
$$\pi^{QLk}_{i,1} = QBR_i(\pi^{QLk}_{-i,0} \mid \lambda_1),$$
$$\pi^{QLk}_{i,2} = QBR_i(\gamma \mid \lambda_2),$$

where γ is a mixed-strategy profile representing level-2 agents' (possibly-incorrect) beliefs about how level-1 agents play, with $\gamma_j(a_j) = QBR_j(\pi^{QLk}_{-j,0} \mid \mu)$. The quantal level-$k$ model thus has five parameters: $\{\alpha_1, \alpha_2\}$, the relative proportions of level-1 and level-2 agents; $\{\lambda_1, \lambda_2\}$, the precisions of level-1 and level-2 agents' responses; and μ, level-2 agents' beliefs about the precision of level-1 agents. ■

Experimental Setup

In this section we describe the data and methods that we used in our model evaluations. We also describe two models based on Nash equilibrium.

Data

During the literature survey described earlier, we also looked for datasets. We identified nine large-scale, publicly-available sets of human-subject experimental data. Of these, five (Stahl and Wilson 1994; 1995; Costa-Gomes, Crawford, and Broseta 1998; Goeree and Holt 2001; Cooper and Van Huyck 2003) were used in follow-up work by researchers other than the original authors; we included all of these datasets in our study. We also included the dataset from (Rogers, Palfrey, and Camerer 2009), as it contained a wide variety of game types, including asymmetric games and games with differing numbers of actions. We excluded the remaining three datasets

[1] We here model only level-k agents, unlike (Costa-Gomes, Crawford, and Broseta 2001) who also modeled other decision rules.

(Haruvy, Stahl, and Wilson 2001), (Haruvy and Stahl 2007), (Stahl and Haruvy 2008) because they were substantially similar to the datasets from (Stahl and Wilson 1994), (Stahl and Wilson 1995), and because of computational resource constraints.

In (Stahl and Wilson 1994) experimental subjects played 10 normal-form games, with payoffs denominated in units worth 2.5 cents. In (Stahl and Wilson 1995), subjects played 12 normal-form games, where each point of payoff gave a 1% chance (per game) of winning $2.00. In (Costa-Gomes, Crawford, and Broseta 1998) subjects played 18 normal-form games, with each point of payoff worth 40 cents. However, subjects were paid based on the outcome of only one randomly-selected game. (Goeree and Holt 2001) presented 10 games in which subjects' behavior was close to that predicted by Nash equilibrium, and 10 other small variations on the same games in which subjects' behavior was *not* well-predicted by Nash equilibrium. Half of these games were normal form; the payoffs for each game were denominated in pennies. In (Cooper and Van Huyck 2003), agents played the normal forms of 8 games, followed by extensive form games with the same induced normal form; we include only the data from the normal-form games. Finally, in (Rogers, Palfrey, and Camerer 2009), subjects played 17 normal-form games, with payoffs denominated in pennies.

We represent each observation of an action by an experimental subject as a pair (a_i, G), where a_i is the action that the subject took when playing as player i in game G. All games were two player, so each single play of a game generated two observations. We built one dataset for each study, named by the source study: SW94 contains 400 observations from (Stahl and Wilson 1994), SW95 has 576 observations from (Stahl and Wilson 1995), CGCB98 has 1566 observations from (Costa-Gomes, Crawford, and Broseta 1998), GH01 has 500 observations from (Goeree and Holt 2001), CVH03 has 2992 observations from (Cooper and Van Huyck 2003), and RPC09 has 1210 observations from (Rogers, Palfrey, and Camerer 2009). We combined the data from all 75 games into a seventh dataset (ALL6) containing 6974 observations.

Methods

To evaluate a given model on a given dataset, we performed 10 rounds of 10-fold cross-validation. Specifically, for each round, we randomly divided the dataset into 10 parts. For each of the 10 ways of selecting 9 parts from the 10, we computed the maximum likelihood estimate of the model's parameters based on those 9 parts, using the Nelder-Mead simplex algorithm (Nelder and Mead 1965). We then determined the log likelihood of the remaining part given the prediction. We call the average of this quantity across all 10 parts the *cross-validated log likelihood*. The average (across rounds) of the cross-validated log likelihoods is distributed according to a Student's-t distribution see, e.g., (Witten and Frank 2000). We compared the predictive power of different behavioral models on a given dataset by comparing the average cross-validated log likelihood of the dataset under each model. We say that one model predicted significantly better than another when the 95% confidence intervals for the average cross-validated log likelihoods do not overlap.

We used GAMBIT (McKelvey, McLennan, and Turocy 2007) to compute QRE and to enumerate the Nash equilibria of games. We performed computation on the glacier cluster of WestGrid (www.westgrid.ca), which consists of 840 computing nodes, each with two 3.06GHz Intel Xeon 32-bit processors and either 2GB or 4GB of RAM. In total, the results reported in this paper required approximately 107 CPU days of machine time, primarily for model fitting.

Nash Equilibrium Models

Any attempt to use Nash equilibrium for prediction must extend the solution concept to solve two problems: ensuring that no action is assigned probability 0, and dealing with multiple equilibria.[2] Indeed, in 83% of the games in the ALL6 dataset (62 out of 75), *every* Nash equilibrium assigned probability 0 to actions that were actually taken by experimental subjects. This means that treating Nash equilibrium as a prediction resulted in the entire dataset having probability 0.

We solved the first problem by adding a parameter representing the probability that a player will choose an action at random. (As in our other models, we fit this parameter using maximum likelihood estimation.) We constructed two new models, corresponding to two ways of solving the second problem. The first model, *uniform Nash equilibrium with error* (UNEE), takes the average over the predictions of every Nash equilibrium. This is equivalent to having a uniform prior over the equilibria of a game; its performance provides a lower bound on the quality of predictions that can be made based on Nash equilibrium. The second model, *nondeterministic Nash equilibrium with error* (NNEE), nondeterministically selects the Nash equilibrium that is most consistent with the full dataset. Clearly this model cannot not be used for prediction, as it relies upon "peeking" at the full dataset. Its performance gives an upper bound on the quality of predictions based on a *single* Nash equilibrium; note, however, that it is possible for UNEE to achieve better performance than NNEE.

Model Analysis

In this section we describe the results of our experiments. Figure 1 compares our four behavioral and two equilibrium-based models. For each model and each dataset, we give the factor by which the dataset is more likely according to the model's prediction than it is according to a uniform random prediction. Thus, for example, the ALL6 dataset is approximately 10^{20} times more likely according to QRE's prediction than it is according to a uniform random prediction.

Comparing Behavioral Models

In most datasets, the model based on cost-proportional errors (QRE) predicted significantly better than the two models based on bounded iterated reasoning (Lk and Poisson-

[2] One might wonder whether the ϵ-equilibrium solution concept solves either of these problems. In fact, it makes the equilibrium selection problem much harder, as every game has infinitely many ϵ-equilibria for any $\epsilon > 0$. To our knowledge, no algorithm for characterizing this set exists, making equilibrium selection impractical. Thus, we did not consider ϵ-equilibrium in our study.

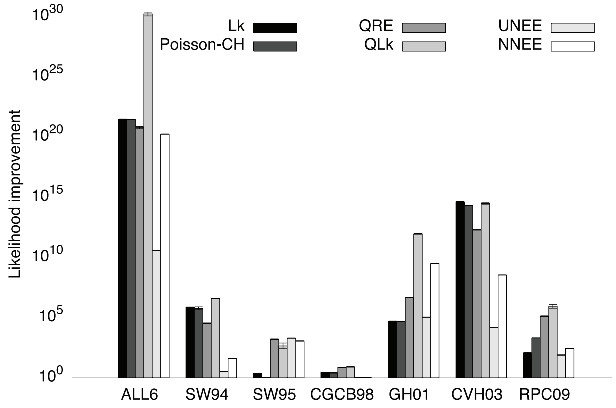


Figure 1: Average likelihood ratios of model predictions to random predictions, with 95% confidence intervals.

CH). However, in three datasets, including the aggregated dataset, the situation was reversed, with Lk and Poisson-CH outperforming QRE. This mixed result is consistent with earlier comparisons of QRE with these two models (Chong, Camerer, and Ho 2005; Crawford and Iriberri 2007; Rogers, Palfrey, and Camerer 2009), and suggests that bounded iterated reasoning and cost-proportional errors capture distinct underlying phenomena. That suggests that our remaining model, which incorporates both components, should predict better than models that incorporate only one component. This was indeed the case, as QLk generally outperformed the single-component models. Overall, QLk was the strongest of the behavioral models, predicting significantly better than all models in all datasets except CVH03 and SW95.

In contrast to earlier studies which found few to no level-0 agents (Stahl and Wilson 1994; 1995; Haruvy, Stahl, and Wilson 2001), our fitted parameters for the Lk and QLk models estimated large proportions of level-0 agents (61% and 54% respectively on the ALL6 dataset). This is explained by differences in the fitting procedures used. We chose parameters to maximize the likelihood of all observed behavior in a given dataset, whereas the cited studies estimated parameters on a per-subject basis by assigning each subject the level that would maximize the likelihood of his or her sequence of choices in all the games in the dataset.

Comparing to Nash Equilibrium

It is already well known that Nash equilibrium is a poor description of humans' initial play in normal-form games e.g., see (Goeree and Holt 2001). However, for the sake of completeness, we also evaluated the predictive power of Nash equilibrium on our datasets. Referring again to Figure 1, we see that UNEE's predictions were significantly worse than those of every behavioral model on every dataset except GH01 and SW95. NNEE's predictions were significantly worse than those of QLk on every dataset except SW95 and GH01. This is strong evidence that behavioral models better predict human play in normal-form games than Nash equilibrium. It is unsurprising that GH01 was an exception, since it was deliberately constructed so that human play on half of its games would be relatively well-described by Nash equilibrium. The performance of UNEE on SW95 is more surprising, and might deserve additional study.

Deeper Analysis of Behavioral Models

In this section we perform a deeper analysis of our four behavioral models in order to answer four questions that arose out of our initial evaluation. Specifically, for each question we constructed modified models and compared their performance to that of the original models. Figure 2 reports the evaluations of all the modified models considered in this section, expressed as a ratio between the likelihood of the modified model and the corresponding original model.

Are Poisson Distributions Helpful in CH?

Our first question was whether it is reasonable to assume that agent levels have a Poisson distribution in the cognitive hierarchy model. At the best-fitting parameter values for ALL6, this would imply that roughly 59% of agents are level-0, which we consider implausible. We hypothesized that a cognitive hierarchy model assuming some other distribution would better fit the data. To test this hypothesis, we constructed a 4-parameter cognitive hierarchy model (CH4), in which each agent was assumed to have level $m \leq 4$, but where the distributional form was otherwise unrestricted.

In Figure 2(a) we can see that the ALL6 dataset is approximately 10,000 times more likely according to the CH4 model's prediction than it is according to Poisson-CH. CH4 predicted significantly better than the Poisson-CH model on most datasets, and never significantly worse. Overall, we conclude that the assumption of Poisson-distributed agent levels was unhelpful in the cognitive hierarchy model.

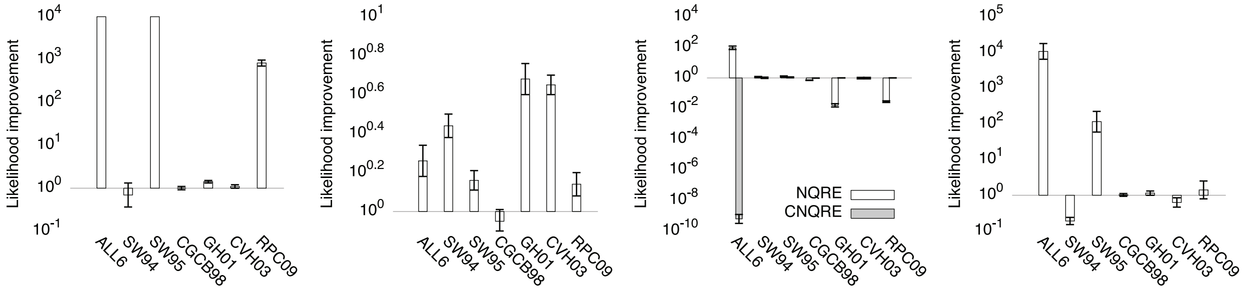
Are Higher Level Agents Helpful in Level-k?

Both the quantal level-k and level-k models assume that all agents have level $k \leq 2$. Our second question was whether a richer model that allowed for higher-level agents would have better predictive power. To explore this question, we constructed a level-k model with $k \in \{0, 1, 2, 3, 4\}$ (Lk4). We hypothesized that the Lk4 model would have better predictive power than the Lk model.

As reported in Figure 2(b), the Lk4 model predicted significantly better than the Lk model on all datasets except CGCB98, where there was no significant difference between the two models. However, these differences were small in every case, in spite of the fact that Lk4 has twice as many parameters as Lk. Overfitting does not appear to have influenced these results, as the ratios of test to training log likelihoods were not significantly different between the Lk and Lk4 models. This suggests that few players in our datasets are well-described as higher-level agents in a level-k model.

Does Payoff Scaling Matter?

Our third question was whether the payoffs in the different games in the dataset were in appropriate units. Unlike the level-k and Poisson-CH models, both QRE and quantal



(a) CH4 vs. Poisson-CH.

(b) Lk4 vs. Lk.

(c) NQRE, CNQRE vs. QRE.

(d) QCH vs. QLk

Figure 2: Average likelihood ratios between predictions of modified and initial models, with 95% confidence intervals.

level-k depend on the units used to represent payoffs in a game. When considering a single setting this is not a concern, because the precision parameter can scale a game to appropriately-sized units. However, when data is combined from multiple studies in which payoffs are expressed on different scales, a single precision parameter may be insufficient to compensate for QRE's scale dependence.

We proposed two hypotheses to explore this question. The first was that subjects were concerned only with relative scales of payoff differences within individual games. To test this hypothesis, we constructed a model (NQRE) that normalizes a game's payoffs to lie in the interval $[0, 1]$, and then predicts based on the QRE of the normalized game. Our second hypothesis was that subjects were concerned with the expected monetary value of their payoffs. To test this hypothesis, we constructed a model (CNQRE) that normalizes payoffs so that they are denominated in expected cents.

Figure 2(c) reports the likelihood ratio between the modified QRE models and QRE. Both NQRE and CNQRE performed worse than the original unnormalized QRE on every disaggregated dataset except for `SW94` and `SW95`, where the improvements were very small (although significant). We conclude that subjects responded to the raw payoff numbers, not to the actual values behind those payoff numbers, and not solely to the relative size of the payoff differences. There are independent reasons to find this plausible, such as the widely-studied "money illusion" effect (Shafir, Diamond, and Tversky 1997), in which people focus on nominal rather than real monetary values.

However, on the aggregated `ALL6` dataset, the situation was quite different, with NQRE performing well and CNQRE performing very poorly. This suggests that normalization can yield a better-performing QRE estimate for aggregated experimental data, but that expected monetary value is not a helpful normalization to use.

Does Heterogeneity Matter?

The quantal level-k model incorporates multiple kinds of heterogeneity. Different agent types may have different quantal choice precisions, and higher-level agents' beliefs about the relative proportions of other levels in the population, as well as the precisions of other levels, may differ from both each other and reality. Our final question was whether a more constrained model would predict equally well.

We constructed a model in which non-random (i.e., non-level-0) agents were constrained to have identical precisions. Further, the agents were constrained to have correct beliefs about the precisions and the relative proportions of lower-level types. This model can also be viewed as an extension of cognitive hierarchy that adds quantal response; hence we called it *quantal cognitive hierarchy*, or QCH.

Definition 6 (QCH)**.** Level-0 agents choose actions uniformly at random. Level-m agents choose actions with probability $\pi^{QCH}_{i,m}(a_i) = QBR_i\left(\sum_{\ell=0}^{m-1} \alpha_\ell \pi^{QCH}_{j,\ell} \mid \lambda\right)$. QCH assumes that $m \leq 4$, and thus has five parameters. ■

Figure 2(d) shows the comparison between the prediction performance of QCH and QLk. QCH actually performed considerably *better* than QLk on the `ALL6` and `SW95` datasets. Otherwise its performance was similar to QLk's, and was never worse by more than a factor of 10. This suggests that QLk's added flexibility in terms of heterogeneous beliefs and precisions did not lead to substantially better predictions.

Conclusions and Overall Recommendations

To our knowledge, this is the first study to address the question of which of the QRE, level-k, cognitive hierarchy, and quantal level-k behavioral models is best suited to predicting unseen human play of normal-form games. We explored the prediction performance of these models, along with several modifications. Overall, we found that the QLk model had substantially better prediction performance than any other model from the literature. We would thus recommend the use of QLk by researchers wanting to predict human play in (unrepeated) normal-form games, especially if maximal accuracy is the main concern. QCH, a novel and conceptually-simpler modification of QLk, performed about as well as QLk. We recommend the use of QCH when it is important to be able to interpret the parameters (e.g., in a Bayesian setting where "reasonable" priors need to be determined) and when it is important to be able to vary the number of modeled levels.

One possible direction for future work is to apply these models to practical applications in multiagent systems. Another is to evaluate models that have been extended to account for learning and non-initial play, including repeated-game and extensive-form game settings.

References

Altman, A.; Bercovici-Boden, A.; and Tennenholtz, M. 2006. Learning in one-shot strategic form games. In *ECML*, 6–17.

Benisch, M.; Davis, G. B.; and Sandholm, T. 2006. Algorithms for rationalizability and CURB sets. In *AAAI*.

Camerer, C.; Ho, T.; and Chong, J. 2001. Behavioral game theory: Thinking, learning, and teaching. Nobel Symposium on Behavioral and Experimental Economics.

Camerer, C.; Ho, T.; and Chong, J. 2004. A cognitive hierarchy model of games. *QJE* 119(3):861–898.

Camerer, C. F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Chong, J.; Camerer, C.; and Ho, T. 2005. Cognitive hierarchy: A limited thinking theory in games. *Experimental Business Research, Vol. III: Marketing, accounting and cognitive perspectives* 203–228.

Conitzer, V., and Sandholm, T. 2005. A generalized strategy eliminability criterion and computational methods for applying it. In *AAAI*, 483–488.

Cooper, D., and Van Huyck, J. 2003. Evidence on the equivalence of the strategic and extensive form representation of games. *JET* 110(2):290–308.

Costa-Gomes, M., and Crawford, V. 2006. Cognition and behavior in two-person guessing games: An experimental study. *AER* 96(5):1737–1768.

Costa-Gomes, M.; Crawford, V.; and Broseta, B. 1998. Cognition and behavior in normal-form games: an experimental study. Discussion paper 98-22, UCSD.

Costa-Gomes, M.; Crawford, V.; and Broseta, B. 2001. Cognition and behavior in normal-form games: An experimental study. *Econometrica* 69(5):1193–1235.

Costa-Gomes, M.; Crawford, V.; and Iriberri, N. 2009. Comparing models of strategic thinking in Van Huyck, Battalio, and Beil's coordination games. *JEEA* 7(2-3):365–376.

Crawford, V., and Iriberri, N. 2007. Fatal attraction: Salience, naivete, and sophistication in experimental "hide-and-seek" games. *AER* 97(5):1731–1750.

Georganas, S.; Healy, P. J.; and Weber, R. 2010. On the persistence of strategic sophistication. Working paper, University of Bonn.

Goeree, J. K., and Holt, C. A. 2001. Ten little treasures of game theory and ten intuitive contradictions. *AER* 91(5):1402–1422.

Hahn, P. R.; Lum, K.; and Mela, C. 2010. A semiparametric model for assessing cognitive hierarchy theories of beauty contest games. Working paper, Duke University.

Halpern, J. Y., and Pass, R. 2009. Iterated regret minimization: A new solution concept. In *IJCAI*, 153–158.

Haruvy, E., and Stahl, D. 2007. Equilibrium selection and bounded rationality in symmetric normal-form games. *JEBO* 62(1):98–119.

Haruvy, E.; Stahl, D.; and Wilson, P. 1999. Evidence for optimistic and pessimistic behavior in normal-form games. *Economics Letters* 63(3):255–259.

Haruvy, E.; Stahl, D.; and Wilson, P. 2001. Modeling and testing for heterogeneity in observed strategic behavior. *Review of Economics and Statistics* 83(1):146–157.

Hyafil, N., and Boutilier, C. 2004. Regret minimizing equilibria and mechanisms for games with strict type uncertainty. In *UAI*, 268–277.

McKelvey, R., and Palfrey, T. 1995. Quantal response equilibria for normal form games. *GEB* 10(1):6–38.

McKelvey, R.; McLennan, A.; and Turocy, T. 2007. Gambit: Software tools for game theory, version 0.2007. 01.30.

Morgan, J., and Sefton, M. 2002. An experimental investigation of unprofitable games. *GEB* 40(1):123–146.

Nelder, J. A., and Mead, R. 1965. A simplex method for function minimization. *Computer Journal* 7(4):308–313.

Rey-Biel, P. 2009. Equilibrium play and best response to (stated) beliefs in normal form games. *GEB* 65(2):572–585.

Rogers, B. W.; Palfrey, T. R.; and Camerer, C. F. 2009. Heterogeneous quantal response equilibrium and cognitive hierarchies. *JET* 144(4):1440–1467.

Shafir, E.; Diamond, P.; and Tversky, A. 1997. Money illusion. *QJE* 112(2):341–374.

Stahl, D., and Haruvy, E. 2008. Level-n bounded rationality and dominated strategies in normal-form games. *JEBO* 66(2):226–232.

Stahl, D., and Wilson, P. 1994. Experimental evidence on players' models of other players. *JEBO* 25(3):309–327.

Stahl, D., and Wilson, P. 1995. On players' models of other players: Theory and experimental evidence. *GEB* 10(1):218–254.

Tennenholtz, M. 2002. Competitive safety analysis: Robust decision-making in multi-agent systems. *JAIR* 17:363–378.

Von Neumann, J., and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton University Press.

Weizsäcker, G. 2003. Ignoring the rationality of others: evidence from experimental normal-form games. *GEB* 44(1):145–171.

Wellman, M.; Greenwald, A.; and Stone, P. 2007. *Autonomous Bidding Agents: Strategies and Lessons from the Trading Agent Competition*. MIT Press.

Witten, I. H., and Frank, E. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.