

Predicting Human Behavior in Unrepeated, Simultaneous-Move Games[☆]

James R. Wright^{1,*}, Kevin Leyton-Brown^{1,**}

^a*Department of Computer Science, 201-2366 Main Mall, University of British Columbia,
Vancouver, BC, Canada, V6T 1Z4*

Abstract

It is common to assume that agents will adopt Nash equilibrium strategies; however, experimental studies have demonstrated that Nash equilibrium is often a poor description of human players' behavior in unrepeated normal-form games. In this paper, we analyze four widely studied models (QRE, Lk, Cognitive Hierarchy, QLk) that aim to describe actual, rather than idealized, human behavior in such games. We performed what we believe is the most comprehensive meta-analysis of these models, leveraging nine different data sets from the literature recording human play of two-player games. We began by evaluating the models' *generalization* or *predictive* performance, asking how well a model fits unseen "test data" after having had its parameters calibrated based on separate "training data". Surprisingly, we found that (what we dub) the QLk model of Stahl and Wilson (1994) consistently achieved the best performance. Motivated by this finding, we describe methods for analyzing the posterior distributions over a model's parameters. We found that QLk's parameters were being set to values that were not consistent with their intended economic interpretations. We thus explored variations of QLk, ultimately identifying a new model family that has fewer parameters, gives rise to more parsimonious

[☆]Preliminary versions of portions of this work appeared in the proceedings of two computer science conferences (Wright and Leyton-Brown, 2010, 2012).

*Principal corresponding author, Telephone +1 604 827 3924

**Corresponding author, Telephone +1 604 822 1453

Email addresses: jrwright@cs.ubc.ca (James R. Wright), kevinlb@cs.ubc.ca (Kevin Leyton-Brown)

parameter values, and achieves better predictive performance.

Keywords: Behavioral game theory, Bounded rationality, Game theory, Cognitive models, Prediction

1. Introduction

In strategic settings, it is common to assume that agents will adopt Nash equilibrium strategies, behaving so that each optimally responds to the others. This solution concept has many appealing properties; e.g., under any other strategy profile, one or more agents will regret their strategy choices. However, experimental evidence shows that Nash equilibrium often fails to describe human strategic behavior (see, e.g., Goeree and Holt, 2001)—even among professional game theorists (Becker et al., 2005). The relatively new field of *behavioral game theory* extends game-theoretic models to account for human behavior by accounting for human cognitive biases and limitations (Camerer, 2003). Experimental evidence is the foundation of behavioral game theory, and researchers have developed many models of how humans behave in strategic situations based on such data. This multitude of models presents a practical problem, however: which model should we use to predict human behavior? Existing work in behavioral game theory does not directly answer this question, for two reasons. First, it has tended to focus on explaining (fitting) in-sample behavior rather than predicting out-of-sample behavior. This means that models are vulnerable to “overfitting” the data: the most flexible model can be chosen instead of the most accurate one. Second, behavioral game theory has tended not to compare multiple behavioral models, instead either exploring elaborations of a single model or comparing only to one other model (typically Nash equilibrium). In this work we perform rigorous—albeit computationally intensive—comparisons of many different models and model variations on a wide range of experimental data, leading us to believe that ours is the most comprehensive study of its kind.

Our focus is on the most basic of strategic interactions: unrepeated (“initial”) play in simultaneous move games. In the behavioral game theory literature, four

key paradigms have emerged for modeling human decision making in this setting: quantal response equilibrium (QRE; McKelvey and Palfrey, 1995); cognitive hierarchy model (CH; Camerer et al., 2004) models; the closely related level- k (Lk; Costa-Gomes et al., 2001; Nagel, 1995) models; and what we dub quantal level- k (QLk; Stahl and Wilson, 1994) models. Although there exist studies exploring different variations of these models (e.g., Stahl and Wilson, 1995; Ho et al., 1998; Rogers et al., 2009), the overwhelming majority of behavioral models of initial play of normal-form games fall broadly into this categorization.

The first main contribution of our work is to conduct an exhaustive meta-analysis based on data published in nine different studies, rigorously comparing Lk, QLk, CH and QRE to each other and to a model based on Nash equilibrium. All of these models depend upon exogenous parameters. Most previous work has focused on models' ability to *describe* human behavior, and hence has sought parameter values that best explain observed experimental data, or more formally that maximize a dataset's probability. (Observe that all of the models that we consider make probabilistic predictions; thus, we must score models according to how much probability mass they assign to observed events, rather than assessing "accuracy.") We depart from this descriptive focus, seeking to find models, and hence parameter values, that are effective for *predicting* previously unseen human behavior. Thus, we follow a different approach taken from machine learning and statistics. We begin by randomly dividing the experimental data into a training set and a test set. We then set each model's parameters to values that maximize the likelihood of the training dataset, and finally score the each model according to the (disjoint) test dataset's likelihood. To reduce the variance of this estimate without biasing its expected value, we employ a process called cross-validation (see, e.g., Bishop, 2006), systematically repeating this procedure with different test and training sets.

Our meta-analysis has led us to draw three qualitative conclusions. First, and least surprisingly, Nash equilibrium is less able to explain human play than behavioral models. Second, two high-level themes that underlie the four behavioral models (which we dub "cost-proportional errors" and "limited iterative

strategic thinking”) appear to model independent phenomena. Third, and building on the previous conclusion, the quantal level- k model of Stahl and Wilson (1994) (QLk)—which combines both of these themes—made the most accurate predictions. Specifically, QLk substantially outperformed all other models on a new dataset spanning all data in our possession, and also had the best or nearly the best performance on each individual dataset. Our findings were quite robust to variation in the games played by human subjects. We broke down model performance by game properties such as dominance structure and number/types of equilibria, and obtained essentially the same results as on the combined dataset. We do note that our datasets consist entirely of two-player games; however, previous work suggests that human subjects reason about n -player games as if they were two-player games (failing to fully account for the independence of the other players’ actions) (Ho et al., 1998; Costa-Gomes et al., 2009); we might thus expect to observe qualitatively similar results in the n -player case.

The approach we have described so far is designed to compare model performance, but yields little insight into how or why a model works. For example, maximum likelihood estimates provide no information about the extent to which parameter values can be changed without a large drop in predictive accuracy, or even about the extent to which individual parameters influence a model’s performance at all. We thus introduce an alternate (Bayesian) approach for gaining understanding about a behavioral model’s entire parameter space. We combine experimental data with explicitly quantified prior beliefs to derive a posterior distribution that assigns probability to parameter settings in proportion to their consistency with the data and the prior (Gill, 2002). Applying this approach, we analyze the posterior distributions for two models: QLk and Poisson-Cognitive Hierarchy (Poisson-CH). Although Poisson-CH did not demonstrate competitive performance in our initial model comparisons, we analyze it because it is very low-dimensional, and because of a very concrete and influential recommendation in the literature: Camerer et al. (2004) recommended setting the model’s single parameter, which represents agents’ mean number of steps of strategic reasoning,

to 1.5. Our own analysis sharply contradicts this recommendation, placing the 99% confidence interval almost a factor of three lower, on the range $[0.51, 0.59]$. We devote most of our attention to QLk, however, due to its extremely strong performance. Our new analysis points out multiple anomalies in QLK’s optimal parameter settings, suggesting that a simpler model could be preferable. We thus exhaustively evaluated a family of variations on QLk, thereby identifying a simpler, more predictive family of models based in part on the cognitive hierarchy concept. In particular, we introduce a new three-parameter model that gives rise to a more plausible posterior distribution over parameter values, while also achieving better predictive performance than (five-parameter) QLk.

In the next section, we define the models that we study. Section 3 lays out the formal framework within which we work, and Section 4 describes our data, methods, and the Nash-equilibrium-based model to which we compare the behavioral models. Section 5 presents the results of our comparisons. Section 6 introduces our methods for Bayesian parameter analysis, and Section 7 describes the anomalies we identified by applying this analysis to our datasets. Section 8 explains the space of QLK variations that we investigated, and introduces our new, high-performing three-parameter model. In Section 9 we survey related work from the literature and explain how our own work contributes to it. We conclude in Section 10.

2. Models for Predicting Human Play of Simultaneous-Move Games

Formally, a behavioral model is a mapping from a game description G and a vector of parameters θ to a predicted distribution over each action profile a in G , which we denote $\Pr(a | G, \theta)$. In what follows, we define four prominent behavioral models of human play in unrepeated, simultaneous-move games.

2.1. Quantal Response Equilibrium

One important idea from behavioral economics is that people become more likely to make errors as those errors become less costly; we call this making *cost-proportional errors*. This can be modeled by assuming that agents best respond *quantally*, rather than via strict maximization.

Definition 1 (Quantal best response). Let $u_i(a_i, s_{-i})$ be agent i 's expected utility in game G when playing action a_i against strategy profile s_{-i} . Then a (logit) *quantal best response* $QBR_i^G(s_{-i}; \lambda)$ by agent i to s_{-i} is a mixed strategy s_i such that

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

where λ (the *precision* parameter) indicates how sensitive agents are to utility differences, with $\lambda = 0$ corresponding to uniform randomization and $\lambda \rightarrow \infty$ corresponding to best response. Note that unlike best response, which is a set-valued function, quantal best response always returns a unique mixed strategy.

The notion of quantal best responses 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, $s_i^* = QBR_i^G(s_{-i}^*; \lambda)$ for all agents i .

A QRE is guaranteed to exist for any normal-form game and non-negative precision (McKelvey and Palfrey, 1995). However, it is not guaranteed to be unique. As is standard in the literature, we select the (unique) QRE that lies on the principal branch of the QRE homotopy at the specified precision. The principal branch has the attractive feature of approaching the risk-dominant equilibrium (as $\lambda \rightarrow \infty$) in 2×2 games with two strict equilibria (Turocy, 2005).

Although Equation (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). The QRE concept nevertheless makes sense if human players are believed to play games differently depending on the magnitudes of the payoffs involved.

2.2. Level- k

Another key idea from behavioral economics is that humans can perform only a limited number of *iterations of strategic reasoning*. The level- k model (Costa-Gomes et al., 2001) captures this idea by associating each agent i with a

level $k_i \in \{0, 1, 2, \dots\}$, 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. Agents are assumed not to 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 $BR_i^G(s_{-i})$ denote the set of i 's best responses in game G to the strategy profile s_{-i} . Let $IBR_{i,k}^G$ denote the *iterative best response set* for a level- k agent i , with $IBR_{i,0}^G = A_i$ and $IBR_{i,k}^G = BR_i^G(IBR_{i,k-1}^G)$. Then the distribution $\pi_{i,k}^{Lk} \in \Pi(A_i)$ that the Lk model predicts for a level- k agent i is defined as

$$\begin{aligned} \pi_{i,0}^{Lk}(a_i) &= |A_i|^{-1}, \\ \pi_{i,k}^{Lk}(a_i) &= \begin{cases} (1 - \epsilon_k)/|IBR_{i,k}^G| & \text{if } a_i \in IBR_{i,k}^G, \\ \epsilon_k/(|A_i| - |IBR_{i,k}^G|) & \text{otherwise.} \end{cases} \end{aligned}$$

The overall predicted distribution of actions is a weighted sum of the distributions for each level:

$$\Pr(a_i | G, \alpha_1, \alpha_2, \epsilon_1, \epsilon_2) = \sum_{\ell=0}^2 \alpha_\ell \cdot \pi_{i,\ell}^{Lk}(a_i),$$

where $\alpha_0 = 1 - \alpha_1 - \alpha_2$. This model thus has 4 parameters: $\{\alpha_1, \alpha_2\}$, the proportions of level-1 and level-2 agents, and $\{\epsilon_1, \epsilon_2\}$, the error probabilities for level-1 and level-2 agents.

2.3. Cognitive Hierarchy

The cognitive hierarchy model (Camerer et al., 2004), like level- k , models agents with heterogeneous bounds on iterated reasoning. It differs from the level- k model in two ways. First, according to this model agents do not make

¹We here model only level- k agents, unlike Costa-Gomes et al. (2001) who also modeled other decision rules.

errors; each agent always best responds to its beliefs. Second, agents of level- m best respond to the full distribution of agents at levels 0 – $(m-1)$, rather than only to level- $(m-1)$ agents. More formally, every agent has an associated level $m \in \{0, 1, 2, \dots\}$. Let f be a probability mass function describing the distribution of the levels in the population. Level- 0 agents play uniformly at random. Level- m agents ($m \geq 1$) best respond to the strategies that would be played in a population described by the truncated probability mass function $f(j | j < m)$.

Camerer et al. (2004) advocate a single-parameter restriction of the cognitive hierarchy model called *Poisson-CH*, in which f is a Poisson distribution.

Definition 4 (Poisson-CH model). Let $\pi_{i,m}^{PCH} \in \Pi(A_i)$ be the distribution over actions predicted for an agent i with level m by the Poisson-CH model. Let $f(m) = \text{Poisson}(m; \tau)$. Let $BR_i^G(s_{-i})$ denote the set of i 's best responses in game G to the strategy profile s_{-i} . Let

$$\pi_{i,0:m}^{PCH} = \sum_{\ell=0}^m f(\ell) \frac{\pi_{i,\ell}^{PCH}}{\sum_{\ell'=0}^m f(\ell')}$$

be the “truncated” distribution over actions predicted for an agent conditional on that agent’s having level $0 \leq \ell \leq m$. Then π^{PCH} is defined as

$$\begin{aligned} \pi_{i,0}^{PCH}(a_i) &= |A_i|^{-1}, \\ \pi_{i,m}^{PCH}(a_i) &= \begin{cases} |BR_i^G(\pi_{i,0:m-1}^{PCH})|^{-1} & \text{if } a_i \in BR_i^G(\pi_{i,0:m-1}^{PCH}); \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The overall predicted distribution of actions is a weighted sum of the distributions for each level,

$$\Pr(a_i | G, \tau) = \sum_{\ell=0}^{\infty} f(\ell) \cdot \pi_{i,\ell}^{PCH}(a_i).$$

The mean of the Poisson distribution, τ , is thus this model’s single parameter.

Rogers et al. (2009) note 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 other iterative models) adequately capture cost-proportional errors, or whether it is beneficial to model cost-proportional errors explicitly.

2.4. 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.² Each agent responds to its beliefs quantally, as in QRE.

A key 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, but have (possibly incorrect) beliefs about the lower-level agents' precision. That is, level-1 and level-2 agents use potentially different precisions (λ 's), and furthermore level-2 agents' beliefs about level-1 agents' precision can be wrong.

Definition 5 (QLk model). The probability distribution $\pi_{i,k}^{QLk} \in \Pi(A_i)$ over actions that QLk predicts for a level- k agent i is

$$\begin{aligned}\pi_{i,0}^{QLk}(a_i) &= |A_i|^{-1}, \\ \pi_{i,1}^{QLk} &= QBR_i^G(\pi_{-i,0}^{QLk}; \lambda_1), \\ \pi_{i,1(2)}^{QLk} &= QBR_i^G(\pi_{-i,0}^{QLk}; \lambda_{1(2)}), \\ \pi_{i,2}^{QLk} &= QBR_i^G(\pi_{i,1(2)}^{QLk}; \lambda_2),\end{aligned}$$

where $\pi_{i,1(2)}^{QLk}$ is a mixed-strategy profile representing level-2 agents' prediction of how other agents will play.³ The overall predicted distribution of actions is the weighted sum of the distributions for each level,

$$\Pr(a_i | G, \alpha_1, \alpha_2, \lambda_1, \lambda_2, \lambda_{1(2)}) = \sum_{k=0}^2 \alpha_k \pi_{i,k}^{QLk}(a_i),$$

where $\alpha_0 = 1 - \alpha_1 - \alpha_2$. The QLk model thus has five parameters: $\{\alpha_1, \alpha_2, \lambda_1, \lambda_2, \lambda_{1(2)}\}$.

²Stahl and Wilson (1994) also consider an extended version of this model that adds a type that plays the equilibrium strategy. In order to avoid the complication of having to specify an equilibrium selection rule, we do not consider this extension (as many of the games in our dataset have multiple equilibria). See Section 4.2 for bounds on the performance of Nash equilibrium predictions on our dataset.

³This can be interpreted either as the level-2 agents' beliefs about the behavior of level-1 agents alone, or it can be understood as modeling level-2 agents' beliefs about both level-1 and level-0 agents, with the presence of additional level-0 agents being captured by a lower precision $\lambda_{1(2)}$. Stahl and Wilson (1994) advocate the latter interpretation.

3. Methods I: Comparing Models

3.1. Prediction Framework

How do we determine whether a behavioral model is well supported by experimental data? An experimental dataset $\mathcal{D} = \{(G_i, a_i) \mid i = 1, 2, \dots, I\}$ is a set of I elements (G_i, a_i) , where G_i is a game and a_i is a (pure) action played by a human player in G_i . (Observe that there is no reason to pair the play of a human player with that of his opponent, as games are unrepeated.) Recall that a behavioral model is a mapping from a game description G_i and a vector of parameters θ to a predicted distribution over each action profile a_i in G_i , which we denote $\Pr(a_i \mid G_i, \theta)$.

Our model can only be used to make predictions when its parameters are instantiated. How should we set these parameters? Our goal is a model that produces accurate probability distributions over the actions of human agents, rather than simply determining the single action most likely to be played. This means that we cannot score different models (or, equivalently, different parameter settings for the same model) using a criterion such as a 0–1 loss function (“accuracy”), which asks how many actions were accurately “predicted”. (For example, the 0-1 loss function evaluates models based purely upon which action is assigned the highest probability, and does not take account of the probabilities assigned to the other actions.) Instead, we evaluate a given model on a given dataset by *likelihood*. That is, we compute the probability of the observed actions according to the distribution over actions predicted by the model. The higher the probability of the actual observations according to the prediction output by a model, the better a prediction we say that the model made. This takes account of the full predicted distribution; in particular, given a particular observed distribution, the prediction that maximizes the likelihood score is the observed distribution itself.⁴

⁴Although the likelihood is what we are interested in conceptually, in practice we operate on the log of the likelihood to avoid precision problems. Since log likelihood is a monotonic function of likelihood, a model that has higher likelihood than another model will always also have higher log likelihood, and vice versa.

Assume that there is some “true” set of parameter values, θ^* , under which the model outputs the true distribution $\Pr(a_i | G_i, \theta^*)$ over action profiles, and that θ^* is independent of G . The maximum likelihood estimate of the parameters based on \mathcal{D} ,

$$\hat{\theta} = \arg \max_{\theta} \Pr(\mathcal{D} | \theta),$$

is an unbiased point estimate of the true set of parameters θ^* , whose variance decreases as I grows. We then use $\hat{\theta}$ to evaluate the model:

$$\Pr(a | G, \mathcal{D}) = \Pr(a | G, \hat{\theta}). \quad (2)$$

The likelihood of a single datapoint $d_i = (G_i, a_i) \in \mathcal{D}$ is

$$\Pr(d_i | \theta) = \Pr(G_i, a_i | \theta).$$

By the chain rule of probabilities, this⁵ is equivalent to

$$\Pr(d_i | \theta) = \Pr(a_i | G_i, \theta) \Pr(G_i | \theta),$$

and by independence of G and θ we have

$$\Pr(d_i | \theta) = \Pr(a_i | G_i, \theta) \Pr(G_i). \quad (3)$$

The datapoints are independent, so the likelihood of the dataset is just the product of the likelihoods of the datapoints,

$$\Pr(\mathcal{D} | \theta) = \prod_{i=1}^I \Pr(a_i | G_i, \theta) \Pr(G_i). \quad (4)$$

The probabilities $\Pr(G_i)$ are constant with respect to θ , and can therefore be disregarded when maximizing the likelihood:

$$\arg \max_{\theta} \Pr(\mathcal{D} | \theta) = \arg \max_{\theta} \prod_{i=1}^I \Pr(a_i | G_i, \theta).$$

⁵To those unfamiliar with Bayesian analysis, quantities such as $\Pr(\mathcal{D})$, $\Pr(G_i)$, and $\Pr(G_i | \theta)$ may seem difficult to interpret or even nonsensical. It is common practice in Bayesian statistics to assign probabilities to any quantity that can vary, such as the games under consideration or the complete dataset that has been observed. Regardless of how they are interpreted, these quantities all turn out to be constant with respect to θ , and so have no influence on the outcome of the analysis.

3.2. Assessing Generalization Performance

Each of the models that we consider depends on parameters that are estimated from the data. In such settings, one must be careful to avoid the problem of “overfitting” the data, where the most flexible model can be preferred to the most accurate model. We do this by estimating parameters on a subset of our dataset (“training data”) and then evaluating the resulting model by computing likelihood scores on the remaining, disjoint “test data”.

Randomly dividing our experimental data into training and test sets introduces variance into the prediction score, since the exact value of the score depends partly upon the random division. To reduce this variance, we perform 10 rounds of 10-fold *cross-validation*. Specifically, for each round, we randomly order the datapoints and then divide them into 10 equal-sized parts. For each of the 10 ways of selecting 9 parts from the 10, we compute the maximum likelihood estimate of the model’s parameters based on those 9 parts. We then determine the likelihood of the remaining part given the prediction. We call the average of this quantity across all 10 parts the *cross-validated likelihood*. The average (across rounds) of the cross-validated likelihoods is distributed according to a Student’s-*t* distribution (see, e.g., Witten and Frank, 2000). We compare the predictive power of different behavioral models on a given dataset by comparing the average cross-validated likelihood of the dataset under each model. We say that one model predicts significantly better than another when the 95% confidence intervals for the average cross-validated likelihoods do not overlap.

4. Experimental Setup

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

4.1. Data

As described in detail in Section 9, we conducted an exhaustive survey of papers that make use of the four behavioral models we consider. We thereby identified nine large-scale, publicly available sets of human-subject experimental

data (Stahl and Wilson, 1994, 1995; Costa-Gomes et al., 1998; Goeree and Holt, 2001; Cooper and Van Huyck, 2003; Rogers et al., 2009; Haruvy et al., 2001; Haruvy and Stahl, 2007; Stahl and Haruvy, 2008). We study all nine of these datasets in this paper, and describe each briefly in what follows.⁶

In Stahl and Wilson (1994) experimental subjects played 10 normal-form games, where every point represented a 1% chance (per game) of winning \$2.50. 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 et al. (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. The payoffs for each game were denominated in pennies. We included the 10 games that were in normal form. In Cooper and Van Huyck (2003), agents played the normal forms of 8 games, followed by extensive form games with the same induced normal forms; we include only the data from the normal-form games. Payoffs were denominated in 10 cent units. In Haruvy et al. (2001), subjects played 15 symmetric 3×3 normal form games. The payoffs were "points" representing a percentage chance of winning \$2.00 for each game. In Haruvy and Stahl (2007), subjects played 20 games, again for payoff points representing a percentage chance of winning \$2.00 per game. Stahl and Haruvy (2008) present new data on 15 games that contain strategies that are dominated in ways that are "obvious" to varying degrees, again for percentage chances of winning \$2.00 per game. Finally, in Rogers et al. (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 (G, a_i) , where a_i is the action that the subject took when playing as

⁶An anonymous referee has brought additional datasets to our attention (Costa-Gomes and Crawford, 2006; Costa-Gomes and Weizsäcker, 2008) of which we were unaware; we plan to include these datasets in future work.

Table 1: Names and contents of each dataset. Units are in expected value.

Source	Games	Observations	Units
Stahl and Wilson (1994)	10	400	\$0.025
Stahl and Wilson (1995)	12	576	\$0.02
Costa-Gomes et al. (1998)	18	1566	\$0.022
Goeree and Holt (2001)	10	500	\$0.01
Cooper and Van Huyck (2003)	8	2992	\$0.10
Rogers et al. (2009)	17	1210	\$0.01
Haruvy et al. (2001)	15	869	\$0.02
Haruvy and Stahl (2007)	20	2940	\$0.02
Stahl and Haruvy (2008)	18	1288	\$0.02
COMB09	128	3600	Normalized

player i in game G . All games had two players, so each single play of a game generated two observations. We built one such dataset for each study, as listed in Table 1. We also constructed a combined dataset, **COMB09**, containing data from all the datasets. The datasets contain very different numbers of observations, ranging from 400 (Stahl and Wilson, 1994) to 2992 (Cooper and Van Huyck, 2003). To prevent **COMB09** from being dominated by the larger datasets, we drew 400 observations uniformly without replacement from each dataset, rather than taking the union of all the observations of the datasets. **COMB09** thus contains 3600 observations.

The QRE and QLk models depend on a precision parameter that is not scale invariant. Specifically, if λ is the correct precision for a game whose payoffs are denominated in cents, then $\lambda/100$ would be the correct precision for a game whose payoffs are denominated in dollars. To ensure consistent estimation of precision parameters, especially in the **COMB09** dataset where observations from multiple studies were combined, we normalized the payoff values for each game to be in expected cents. As described earlier, in some datasets, payoff points were worth a certain number of cents; in others, points represented percentage chances of winning a certain sum, or were otherwise in “expected” units. Table 1 lists the number of expected cents that we deemed each payoff point to be worth for the purposes of normalization.

4.2. Comparing to Nash Equilibrium

It is desirable to compare the predictive performance of our behavioral models to that of Nash equilibrium. However, such a comparison is not as simple as one might hope, because any attempt to use Nash equilibrium for prediction must extend the solution concept to address two problems. The first problem is that many games have multiple Nash equilibria; in these cases, the Nash “prediction” is not well defined. The second problem is that Nash equilibrium frequently assigns probability zero to some actions. (Indeed, in 72% of the games in our COMBO9 dataset *every* Nash equilibrium assigned probability 0 to actions that were actually taken by one or more experimental subjects.) This is a problem because we assess the quality of a model by how well it explains the data; unmodified, Nash equilibrium model considers our experimental data to be *impossible*, and hence receives a likelihood of zero.⁷

We addressed the second problem by augmenting the Nash equilibrium solution concept to say that with some probability, each player chooses an action uniformly at random; this prevents the solution concept from assessing any experimental data as impossible. This probability is a free parameter of the model; as we did with behavioral models, we fit this parameter using maximum likelihood estimation on a training set. (We thus call the model Nash Equilibrium with Error, or NEE.) We sidestepped the first problem by assuming that agents always coordinate to play an equilibrium and reporting statistics across different equilibria. Specifically, we report the performance achieved by choosing the equilibria that respectively best and worst fit the *test* data, thereby giving upper and lower bounds on the test-set performance achievable by any Nash-based prediction. (Note that because we “cheat” by choosing equilibria based on test-set performance, this model is not able to generalize to new data, and hence cannot be used in practice.) Finally, we also report the *expected* prediction

⁷One might wonder whether the ϵ -equilibrium solution concept solves either of these problems. It does not. First, relaxing the equilibrium concept only increases the number of equilibria; indeed, every game has infinitely many ϵ -equilibria for any $\epsilon > 0$. Furthermore, to our knowledge, no algorithm for characterizing this set exists, making equilibrium selection impractical. Second, ϵ -equilibrium can still assign probability 0 to some actions.

performance that is achieved by sampling a Nash equilibrium uniformly at random and assuming that agents play this equilibrium; this model *can* be evaluated without looking at the test set, and hence can be used in practice.

4.3. Computational Environment

We performed computation on the `glacier`, `hermes`, and `orcinus` clusters of WestGrid (www.westgrid.ca), which have 1680 32-bit Intel Xeon CPU cores, 672 64-bit Intel Xeon CPU cores, and 9600 64-bit Intel Xeon CPU cores, respectively. In total, computing the results reported in this paper required over 400 CPU-days of machine time, primarily for model fitting and posterior estimation (which is described in Section 6). Specifically, we used GAMBIT (McKelvey et al., 2007) to compute QRE and to enumerate the Nash equilibria of games, and computed maximum likelihood estimates using the Nelder–Mead simplex algorithm (Nelder and Mead, 1965).

5. Model Comparisons

In this section we describe the results of our experiments comparing the predictive performance of the four behavioral models from Section 2 and of the Nash-based models of Section 4.2. Figure 1 compares our behavioral and Nash-based models. For each model and each dataset, we give the factor by which the dataset was judged more likely according to the model’s prediction than it was according to a uniform random prediction. Thus, for example, the COMBO9 dataset was found to be approximately 10^{18} times more likely according to Poisson-CH’s prediction than according to a uniform random prediction. For the Nash Equilibrium with Error model, the error bars show the upper and lower bounds on predictive performance obtained by selecting an equilibrium to maximize or minimize test-set performance, and the main bar shows the expected predictive performance of selecting an equilibrium uniformly at random. For other models, the error bars indicate 95% confidence intervals; in most cases, these intervals are imperceptibly narrow.

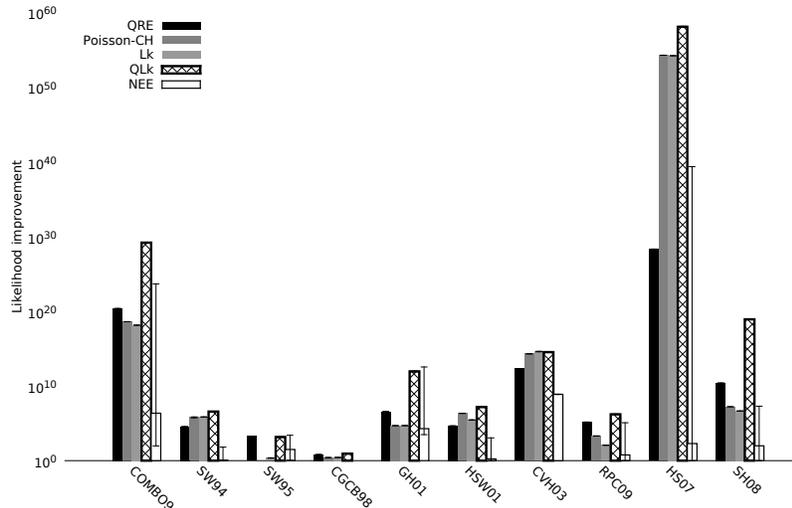


Figure 1: Average likelihood ratios of model predictions to random predictions, with 95% confidence intervals. Error bars for NEE show upper and lower bounds on performance depending upon equilibrium selection; the main bar for NEE shows the average performance over all equilibria.

5.1. Comparing Behavioral Models

Poisson-CH and Lk had very similar performance in most datasets. We did not expect such a strong similarity, given the various differences between the models, but nevertheless found the result intuitive, since the models are very similar to each other. In six datasets, including COMB09, the model based on cost-proportional errors (QRE) predicted human play significantly better than the two models based on bounded iterated reasoning (Lk and Poisson-CH). However, in four datasets, 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 et al., 2005; Crawford and Iriberry, 2007a; Rogers et al., 2009), and suggests to us that, in answer to the question posed in Section 2.3, there may be value to modeling both bounded iterated reasoning and cost-proportional errors explicitly. If this hypothesis is true, we might expect that our remaining model, which incorporates both components, would predict better than models that incorporate only one component. This was indeed the case: QLk generally outperformed the single-component models. Overall, QLk

was the strongest behavioral model, predicting significantly best in all datasets except CVH03 and SW95 (and GH01, which we discuss in detail below).

5.1.1. Level-0

Earlier studies found support for quite variable proportions of level-0 agents. Stahl and Wilson (1994) estimated 0% of the population was level-0;⁸ Stahl and Wilson (1995) estimated 17%, with a confidence interval of [6%, 30%]; Haruvy et al. (2001) estimated rates between 6–16% for various model specifications; and Burchardi and Penczynski (2011) estimated 37% by fitting a level- k model, and between 20–42% by eliciting subject strategies. Our fitted parameters for the Lk and QLk models estimate proportions of level-0 agents that are at the high end of this range (56% and 38% respectively on the COMB09 dataset). However, note that our estimate for QLk is nearly in the center of the range that Burchardi and Penczynski (2011) estimated by directly evaluating subjects’ elicited strategies in a single game. We analyze the full distributions of parameter values in Section 7.

In contrast to our estimates, the number of level-0 agents in the population is typically assumed to be negligible in studies that use an iterative model of behavior. Indeed, some studies (e.g., Crawford and Iriberri, 2007b) fix the number of level-0 agents to be 0. Thus, one possible interpretation of our higher estimates of level-0 agents is as evidence of a misspecified model. For example, Poisson-CH uses level-0 agents as the only source of noisy responses. However, we estimated substantial proportions of level-0 agents even for models (Lk and QLk) that include explicit error structures. We thus believe that the alternative—the possibility that these results point to a substantial frequency of nonstrategic behavior—must be taken seriously.

5.2. Comparing to Nash Equilibrium

It is already widely believed that Nash equilibrium is a poor description of humans’ initial play in normal-form games (e.g., see Goeree and Holt, 2001). Nevertheless, for the sake of completeness, we also evaluated the predictive power

⁸Their dataset is an outlier in our own per-dataset parameter fits; see Section 7.1.

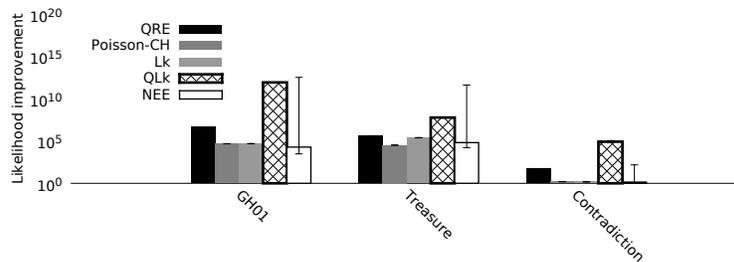


Figure 2: Average likelihood ratios of model predictions to random predictions, with 95% confidence intervals, on **GH01** data separated into “treasure” and “contradiction” treatments. Error bars for NEE show upper and lower bounds on performance depending upon equilibrium selection; the main bar for NEE shows the average performance over all equilibria. Note that relative differences in likelihood are not meaningful across datasets (likelihood drops with growth in the dataset’s number of samples and underlying games’ numbers of actions). Relative differences in likelihood *are* meaningful within datasets.

of Nash equilibrium with error on our datasets. Referring again to Figure 1, we see that NEE’s predictions were worse than those of every behavioral model on every dataset except **SW95**. NEE’s predictions were significantly worse than those of QLk on every dataset except **SW95** and **GH01**.

NEE’s strong performance on **SW95** was surprising; it may have been a result of the unusual subject pool, which consisted of fourth and fifth year undergraduate finance and accounting majors. In contrast, it is unsurprising that NEE performed well on **GH01**, since this distribution was deliberately constructed so that human play on half of its games (the “treasure” conditions) would be relatively well described by Nash equilibrium. Figure 2 separates **GH01** into its “treasure” and “contradiction” treatments and compares the performance of the behavioral and Nash-based models on these separated datasets. In addition to the fact that the “treasure” games were deliberately selected to favor Nash predictions, many of **GH01**’s games have multiple equilibria. This offers an advantage to our NEE model’s upper bound, because it gets to pick the equilibrium with best test-set performance on a per-instance basis (see Section 5.3 below). Note that although NEE thus had a higher upper bound than QLk on the “treasure” treatment, its expected performance was still quite poor.

NEE has a free parameter, ϵ , that describes the probability of an agent

choosing an action uniformly at random. If Nash equilibrium were a good tool for predicting human behavior, we would expect to find this parameter set to a low value; in contrast, the values of ϵ that maximize NEE’s performance were extremely high. On the COMBO9 dataset, a value of $\epsilon = 0.82$ maximized NEE’s average-case performance. Even best-case performance, which is computed by choosing the post-hoc performance-maximizing equilibrium for each game, was optimized by $\epsilon = 0.61$. This does not correspond to our intuition about why playing (part of) a Nash equilibrium would be justified. Thus, the fact that well over half of NEE’s prediction consists of the uniform noise term provides a strong argument against using Nash equilibrium to predict initial play.

5.3. Dataset Composition

As we have already seen in the case of GH01, model performance is sensitive to choices made by the authors of our various datasets about which games to study. One way to control for such choices is to partition our set of games according to important game properties (particularly, dominance solvability and number of equilibria), and to evaluate model performance in each partition. In this section we describe such an analysis.

Overall, our datasets spanned 128 games. The vast majority of these games are matrix games, deliberately lacking inherent meaning in order to avoid framing effects. (Indeed, some studies (e.g., Rogers et al., 2009) even avoid “focal” payoffs like 0 and 100.) For the most part, these games were chosen to vary according to dominance solvability and equilibrium structure. In particular, most dataset authors were concerned with (1) whether a game could be solved by iterated removal of dominated strategies (either strict or weak) and with how many steps of iteration were required; and (2) the number and type of Nash equilibria that each game possesses.⁹

⁹There were two exceptions. The first was Goeree and Holt (2001), who chose games that had both equilibria that human subjects find intuitive and strategically equivalent variations of these games whose equilibria human subjects find counterintuitive. The second exception was Cooper and Van Huyck (2003), whose normal form games were based on an exhaustive enumeration of the payoff orderings possible in generic 2-player, 2-action extensive-form games.

Table 2: Datasets conditioned on various game features. The column headed “games” indicates how many games of the full dataset meet the criterion, and the column headed “ n ” indicates how many observations each feature-based dataset contains. Observe that the game features are not all mutually exclusive, and so the “games” column does not sum to 128.

Name	Description	Games	n
D1	Weak dominance solvable in one round	2	748
D2	Weak dominance solvable in two rounds	38	5058
D2s	Strict dominance solvable in two rounds	23	2000
DS	Weak dominance solvable	44	5446
DSs	Strict dominance solvable	28	2338
ND	Not weak dominance solvable	84	6625
PSNE1	Single Nash equilibrium, which is pure	42	4431
MSNE1	Single Nash equilibrium, which is mixed	24	2509
Multi-Eqm	Multiple Nash equilibria	62	5131

We thus constructed subsets of the full dataset based on their dominance solvability and the nature of their Nash equilibria, as described in Table 2. We used the full dataset with no subsampling rather than `Combo9` because we had less concern about one study dominating a dataset that had been filtered to contain games of a specific type. We computed cross-validated MLE fits for each model on each of the feature-based datasets of Table 2. The results are summarized in Figure 3. In two respects, the results across the feature-based datasets mirror the results of Section 5.1 and Section 5.2. First, QLk significantly outperformed the other behavioral models on almost every dataset; the sole exception is D1, where QLk performed insignificantly better than Lk. Second, every behavioral model significantly outperformed NEE in all but three datasets: D1, ND and Multi-eqm. In these three datasets, the upper and lower bounds on NEE’s performance contained the performance of either two or all three of the single-factor behavioral models (but not QLk). The performance of all models on the D1 dataset was roughly similar, likely due to the ease with which various different forms of reasoning can uncover a dominant strategy. It is unsurprising that NEE’s upper and lower bounds were widely separated on the Multi-eqm dataset, since the more equilibria a game has, the more variation there can be in these equilibria’s post-hoc performance; NEE’s strong best-case performance

on this dataset should similarly reflect this variation. It turns out that 55 of the 84 games (and 4731 of the 6625 observations) in the ND dataset are from the `Multi-eqm` dataset, which likely explains NEE’s high upper bound in that dataset as well. Indeed, this analysis helps to explain some of our previous observations about the `GH01` dataset. NEE contains all other models in its performance bounds in this dataset, and in addition to the fact that half the dataset’s games (the “treasure” treatments) that were chosen for consistency with Nash equilibrium, some of the other games (the “contradiction” treatments) turn out to have multiple equilibria. Overall, the overlap between `GH01` and `Multi-eqm` is 5 games out of 10, and 250 observations out of 500.

Unlike in the per-dataset comparisons of Section 5.1, both of our iterative single-factor models (`Poisson-CH` and `Lk`) significantly outperformed QRE in almost every feature-based dataset, with `D2S` as the only exception. One possible explanation is that the filtering features are all biased toward iterative models. However, it seems unlikely that, e.g., *both* dominance-solvability and dominance-nonsolvability are biased toward iterative models. Another possibility is that iterative models are a better model of human behavior, but the cost-proportional error model of QRE is sufficiently superior to the respectively simple and non-existent error models of `Poisson-CH` and `Lk` that it outperforms on many datasets that mix game types. Or, similarly, iterative models may fit very differently on dominance solvable and non-dominance solvable games; in this case, they would perform very poorly on mixed data. We explore this last possibility in more detail in Section 7.1.

6. Methods II: Analyzing Model Parameters

Making good predictions from behavioral models depends upon obtaining good estimates of model parameters. However, these estimates can also be useful in themselves, helping researchers to understand both how people behave in strategic situations and whether a model’s behavior aligns or clashes with its intended economic interpretation. Unfortunately, the method we have used so far—maximum likelihood estimation, i.e., finding a single set of parameters

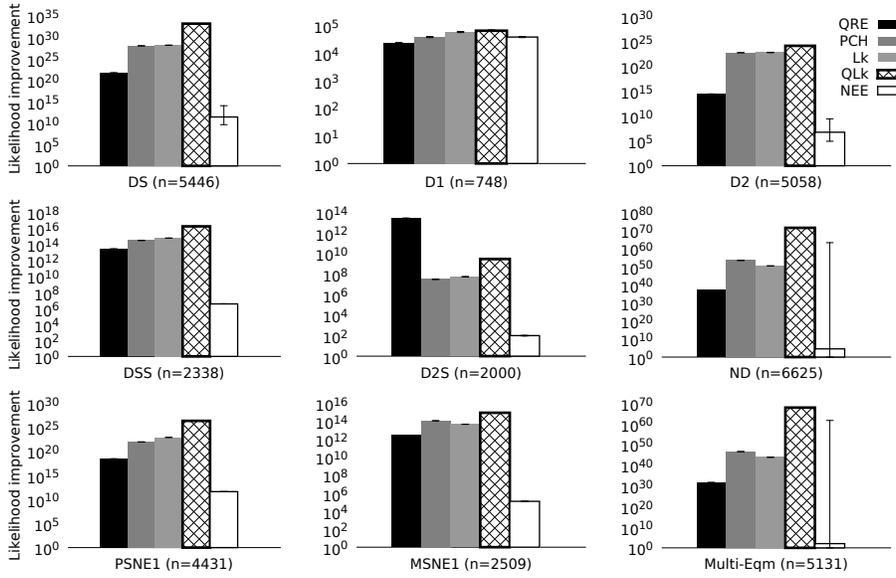


Figure 3: Average likelihood ratios of model predictions to random predictions, with 95% confidence intervals, on feature-based datasets. For NEE the main bar shows performance averaged over all equilibria, and error bars show post-hoc upper and lower bounds on equilibrium performance.

that best explains the training set—is not a good way of gaining this kind of understanding. The problem is that we have no way of knowing how much of a difference it would have made to have set the parameters differently, and hence how important each parameter setting is to the model’s performance. For example, if some parameter is completely uncorrelated with predictive accuracy, the maximum likelihood estimate will set it to an arbitrary value, from which we would be wrong to draw economic conclusions.¹⁰ Similarly, if there are multiple, very different ways of configuring the model to make good predictions, we would not want to draw firm conclusions about how people reason based on a single configuration.

An alternative is to use Bayesian analysis to estimate the entire posterior

¹⁰We can gain local information about the a parameter’s importance from the confidence interval around its maximum likelihood estimate: locally important parameters will have narrow confidence intervals, and locally irrelevant parameters will have wide confidence intervals. However, this does not tell us anything outside the neighborhood of the estimate.

distribution over parameter values, rather than estimating only a mode of this distribution. This allows us to identify the most likely parameter values; how wide a range of values are argued for by the data (equivalently, how strongly the data argues for the most likely values); and whether the values that the data argues for are plausible in terms of our intuitions about parameters’ meanings. In this section we derive an expression for the posterior distribution, and describe methods for constructing posterior estimates and using them to assess parameter importance. In Section 7 we will apply these methods to study QLk and Poisson-CH: the former because it achieved such reliably strong performance, and the latter because it is the model about which the most explicit parameter recommendation was made in the literature.

6.1. Posterior Distribution Derivation

We derive an expression for the posterior distribution $\Pr(\theta | \mathcal{D})$ by applying Bayes’ rule, where $p_0(\theta)$ is the prior distribution:

$$\Pr(\theta | \mathcal{D}) = \frac{p_0(\theta) \Pr(\mathcal{D} | \theta)}{\Pr(\mathcal{D})}. \quad (5)$$

Substituting in Equation (4), which gave an expression for the likelihood of the dataset $\Pr(\mathcal{D} | \theta)$, we obtain

$$\Pr(\theta | \mathcal{D}) = \frac{p_0(\theta) \prod_{i=1}^I \Pr(a_i | G_i, \theta) \Pr(G_i)}{\Pr(\mathcal{D})}. \quad (6)$$

In practice $\Pr(G_i)$ and $\Pr(\mathcal{D})$ are constants, and so can be ignored:

$$\Pr(\theta | \mathcal{D}) \propto p_0(\theta) \prod_{i=1}^I \Pr(a_i | G_i, \theta). \quad (7)$$

Note that by commutativity of multiplication, this is equivalent to performing iterative Bayesian updates one datapoint at a time. Therefore, iteratively updating this posterior neither over- nor underprivileges later datapoints.

6.2. Posterior Distribution Estimation

We estimate the posterior distribution as a set of samples. When a model has a low-dimensional parameter space, like Poisson-CH, we generate a large

number of evenly-spaced, discrete points (so-called *grid sampling*). This has the advantage that we are guaranteed to cover the whole space, and hence will not miss large, important regions. However, this approach does not work when a model’s parameter space is large, because evenly-spaced grids require an number of samples exponential in the number of parameters. Luckily, we do not care about having good estimates of the whole posterior distribution—what matters is getting good estimates of regions of high probability mass. This can be achieved by sampling parameter settings in proportion to their likelihood, rather than uniformly. A wide variety of techniques exist for performing this sort of sampling; we investigated extensively, and achieved the most success with a sequential Monte Carlo technique called *annealed importance sampling*, or AIS (Neal, 2001). AIS allows for efficient sampling from high dimensional distributions, like Markov Chain Monte Carlo (MCMC) techniques. However, each sample point generated using AIS is independent, meaning that AIS does not exhibit the random-walk behavior that can plague MCMC samplers.

Briefly, the annealed importance sampling procedure is as follows. A sample $\vec{\theta}_0$ is drawn from an easy-to-sample-from distribution P_0 . For each P_j in a sequence of intermediate distributions P_1, \dots, P_{r-1} that become progressively closer to the posterior distribution, a sample $\vec{\theta}_j$ is generated by drawing a sample $\vec{\theta}'$ from a *proposal distribution* $Q(\cdot | \vec{\theta}_{j-1})$, and accepted with probability

$$\frac{P_j(\vec{\theta}')Q(\vec{\theta}_{j-1} | \vec{\theta}')}{P_j(\vec{\theta}_{j-1})Q(\vec{\theta}' | \vec{\theta}_{j-1})}. \quad (8)$$

If the proposal is accepted, $\vec{\theta}_j = \vec{\theta}'$; otherwise, $\vec{\theta}_j = \vec{\theta}_{j-1}$. We repeat this procedure multiple times, obtaining one sample each time. In the end, our estimate of the posterior is the set of $\vec{\theta}_r$ values, each weighted according to

$$\frac{P_1(\vec{\theta}_0)P_2(\vec{\theta}_1) \dots P_{r-1}(\vec{\theta}_{r-2})P_r(\vec{\theta}_{r-1})}{P_0(\vec{\theta}_0)P_1(\vec{\theta}_1) \dots P_{r-2}(\vec{\theta}_{r-2})P_{r-1}(\vec{\theta}_{r-1})}. \quad (9)$$

We use a flat prior for all parameters.¹¹ Although this prior is improper on unbounded parameters such as precision, it results in a correctly normalized

¹¹For precision parameters, a flat prior on the log of precision would have been a more

posterior distribution¹²; the posterior distribution in this case reduces to the likelihood (Gill, 2002). For Poisson-CH, where we grid sample an unbounded parameter, we grid sampled within a bounded range $([0, 10])$, which is equivalent to assigning probability 0 to points outside the bounds. In practice, this turns out not to matter, as the vast majority of probability mass was concentrated near 0.

6.3. Visualizing Multi-Dimensional Distributions

In the sections that follow, we present posterior distributions as cumulative marginal distributions. That is, for every parameter, we plot the cumulative density function (CDF)—the probability that the parameter should be set less than or equal to a given value—averaging over values of all other parameters. Plotting cumulative density functions allows us to visualize an entire continuous distribution without having to estimate density from discrete samples, thus sparing us manual decisions such as the width of bins for a histogram. Plotting marginal distributions allows us to examine intuitive two-dimensional plots about multi-dimensional distributions. Interaction effects between parameters are thus obscured; luckily, in separate tests we have found that for our data these were not a major factor.

7. Bayesian Analysis of Model Parameters

In this section we analyze the posterior distributions of the parameters for two of the models compared in Section 5: Poisson-CH and QLk. For Poisson-CH, we computed the likelihood for each value of $\tau \in \{0.01k \mid k \in \mathbb{N}, 0 \leq 0.01k \leq 10\}$, and then normalized by the sum of the likelihoods. For QLk, we used annealed importance sampling. For the initial sampling distribution P_0 , we used a product distribution over the population proportion parameters and the precision

standard choice. In this work, we wanted to avoid artificially preferring precision estimates closer to zero, since it is common for iterative models to assume agents best respond nearly perfectly to lower levels.

¹²That is, for the posterior, $\int \cdots \int_{-\infty}^{\infty} \Pr(\theta \mid \mathcal{D}) d\vec{\theta} = 1$ even though for the prior $\int \cdots \int_{-\infty}^{\infty} p_0(\theta) d\vec{\theta}$ diverges.

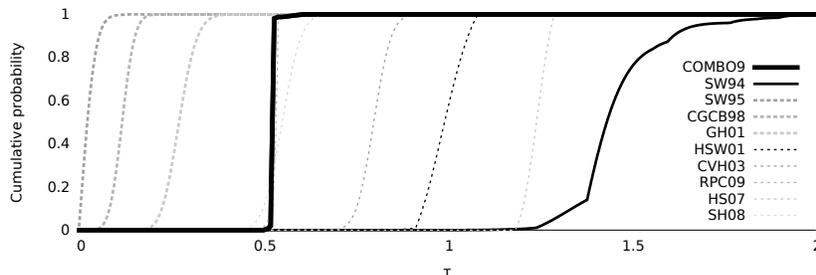


Figure 4: Cumulative posterior distributions for Poisson-CH’s τ parameter. Bold trace is the combined dataset; solid trace is the outlier Stahl and Wilson (1994) source dataset; dotted traces are all other source datasets.

parameters. For the population proportion parameter components we used a Dirichlet distribution $\text{Dir}(1, 1, 1)$; this is equivalent to uniformly sampling over the simplex of all possible combinations of population proportions. For the precision parameter components we used the renormalized non-negative half of a univariate Gaussian distribution $\mathcal{N}(0, 2^2)$ for each precision parameter; this gives a distribution that is decreasing in precision (on the assumption that higher precisions are less likely than lower ones), and with a standard deviation of 2, which was large enough to give a non-negligible probability to most precision estimates from the literature. For the proposal distribution, we chose a product distribution “centered” at the current value, with proportion parameters $\vec{\alpha}'$ sampled from $\text{Dir}(20\vec{\alpha}_{j-1})$, and each precision parameter λ' sampled from $\mathcal{N}(\lambda_{j-1}, 0.2^2)$ (truncated at 0 and renormalized). We chose the “hyperparameters” for the Dirichlet distribution (20) and the precision distributions (0.2^2) by trial and error on a small subset of the data to bring the acceptance rate close to the standard heuristic value of 0.5 (Robert and Casella, 2004). We used 200 intermediate distributions of the form $P_j(\vec{\theta}) = \Pr(\vec{\theta} | \mathcal{D})^{\gamma_j}$, with the first 40 γ_j ’s spaced uniformly from 0 to 0.01, and the remaining 160 γ_j ’s spaced geometrically from 0.01 to 1, as in the original AIS description (Neal, 2001). We performed 5 Metropolis updates in each distribution before moving to the next distribution in the chain.

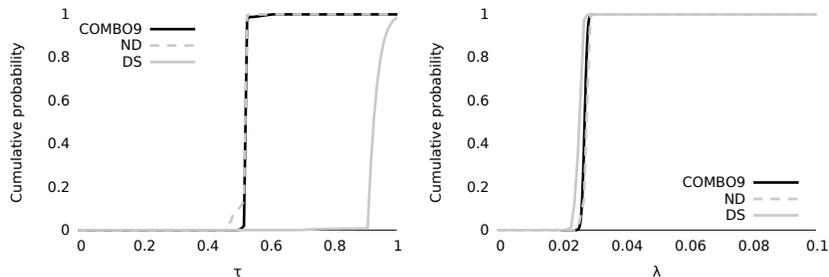


Figure 5: Distributions for the Poisson-CH and QRE models on the combined dataset (COMBO9), non-dominance-solvable games only (ND), and dominance-solvable games only (DS). Left: cumulative distribution of τ parameter of Poisson-CH; Right: cumulative distribution of λ parameter of QRE.

7.1. Poisson-CH

In an influential recommendation from the literature, Camerer et al. (2004) suggest¹³ setting the τ parameter of the Poisson-CH model to 1.5. Our Bayesian analysis techniques allow us to estimate CDFs for this parameter on each of our datasets (see Figure 4). Overall, our analysis strongly contradicts Camerer et al.’s recommendation. On COMBO9, the posterior probability of $0.51 \leq \tau \leq 0.59$ is more than 99%. Every other source dataset had a wider 99% *credible interval* (the Bayesian counterpart to confidence intervals) for τ than COMBO9, as indicated by the higher slope of COMBO9’s cumulative density function (since smaller datasets lead to less confident predictions). Nevertheless, all but two of the source datasets had median values less than 1.0. Only the Stahl and Wilson (1994) dataset (SW94) supports Camerer et al.’s recommendation (median 1.43). However (as we have observed before), SW94 appears to be an outlier; its credible interval is wider than that of the other distributions, and the distribution is very multimodal, likely due to the dataset’s small size.

As we speculated in Section 5.3, Poisson-CH does indeed appear to treat dominance-solvable games differently from non-dominance-solvable games. The left panel of Figure 5 compares the cumulative distribution for τ on the combined

¹³ Although Camerer et al. phrase their recommendation as a reasonable “omnibus guess,” it is often cited as an authoritative finding (e.g., see Carvalho and Santos-Pinto, 2010; Frey and Goldstone, 2011; Choi, 2012; Goodie et al., 2012).

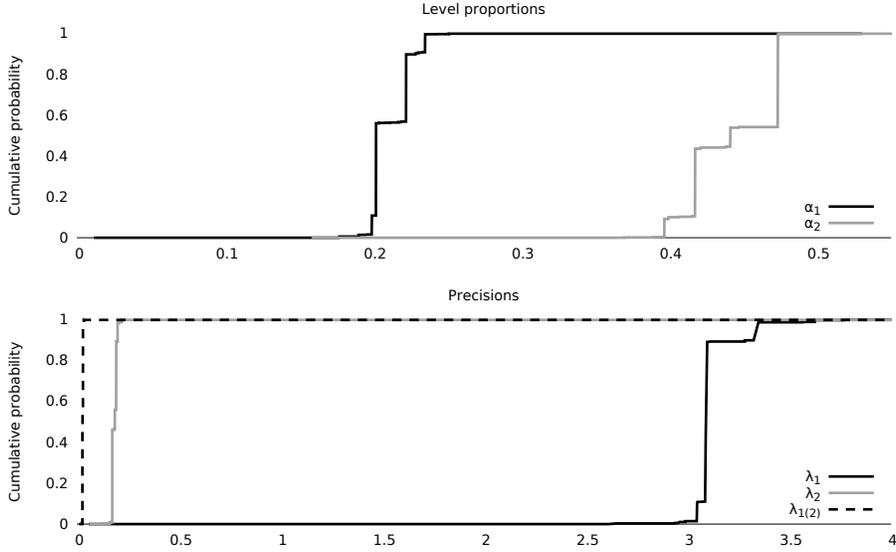


Figure 6: Marginal cumulative posterior distribution functions for the level proportion parameters (α_1, α_2 ; top panel) and precision parameters ($\lambda_1, \lambda_2, \lambda_{1(2)}$; bottom panel) of the QLk model on the combined dataset.

dataset to CDFs for non-dominance-solvable games only and for dominance-solvable games only. The τ parameter has a nearly identical credible interval for non-dominance-solvable games as for the full combined dataset. In contrast, for dominance-solvable games the τ parameter's 99% credible interval is $[0.91, 1.01]$. In contrast, the fits for QRE's precision (λ) parameter are nearly identical for the three game types. In fact, the best-fitting precision for dominance-solvable games is actually slightly smaller, leading to a higher-entropy prediction. This explains how iterative models could outperform QRE on every feature-based dataset in Section 5.3, in spite of being frequently outperformed by QRE in Section 5.1: the features used to separate games in Section 5.3 tend to cluster dominance-solvable and non-dominance-solvable games, and the iterative models can adapt their predictions accordingly, whereas QRE's predictions are not influenced by a game's dominance solvability or lack thereof.

7.2. QLk

Figure 6 gives the marginal cumulative posterior distributions for each of the parameters of the QLk model. (That is, we computed the five-dimensional posterior distribution, and then extracted from it the five marginal distributions shown here.) We found these distributions surprising for several reasons. First, the models predict many more level-2 agents than level-1 agents. In contrast, it is typically assumed that higher level agents are scarcer, as they perform more complex strategic reasoning. Even more surprisingly, the model predicts that level-1 agents should have much higher precisions than level-2 agents. This is odd if the level-2 agents are to be understood as “more rational”; indeed, precision is sometimes interpreted as a measure of rationality (e.g., see Weizsäcker, 2003; Gao and Pfeffer, 2010). Aside from the relative ordering, the median value of λ_1 (3.1) is nearly 20 times larger than that of λ_2 (0.18). It seems unlikely that level-1 agents would be an order of magnitude more sensitive to utility differences than level-2 agents. Finally, the distribution of $\lambda_{1(2)}$ is very concentrated around very small values ([0.023, 0.034]). This is anomalous if $\lambda_{1(2)}$ is interpreted as the precision that level-2 agents ascribe to level-1 agents, since it differs by two orders of magnitude from the “true” value of λ_1 , which is quite concentrated around its median value of 3.1. Alternatively, $\lambda_{1(2)}$ ’s value may simply indicate that level-2 agents believe that a large fraction of the population is level-0.

One possibility is that the QLk model is essentially accurate, and these parameter values simply reflect a surprising reality. For example, the low precision of level-2 agents (and of their predictions of other agents’ behavior) may indicate that two-level strategic reasoning causes a high cognitive load, which makes agents more likely to make mistakes, both in their own actions and in their predictions. The main appeal of this explanation is that it allows us to accept the QLk model’s strong performance at face value. Alternately, we might worry that QLk fails to capture some crucial aspect of experimental subjects’ strategic reasoning. For example, if the low value of $\lambda_{1(2)}$ reflects level-2 agents’ reasoning about all lower levels rather than just one level below themselves, then a model that explicitly represents this process might be preferable to one

that does so via precision-based approximations.¹⁴ We investigate this second possibility in the next section.

8. Model Variations

In this section, we investigate the properties of the QLk model by evaluating the predictive power of a family of systematic variations of the model. In the end, we identify a simpler model that dominated QLk on our data, and which also yielded much more reasonable marginal distributions over parameter values.

Specifically, we constructed a broad family of models by modifying the QLk model along four different axes. First, QLk assumes a maximum level of 2; we considered maximum levels of 1 and 3 as well. Second, QLk assumes *inhomogeneous precisions* in that it allows each level to have a different precision; we varied this by also considering *homogeneous precision* models. Third, QLk allows *general precision beliefs* that can differ from lower-level agents’ true precisions; we also constructed models that make the simplifying assumption that all agents have *accurate precision beliefs* about lower-level agents.¹⁵ Finally, in addition to *Lk* beliefs (where all other agents are assumed by a level-*k* agent to be level- $(k - 1)$), we also constructed models with *CH* beliefs (where agents believe that the population consists of the true, truncated distribution over the lower levels). We evaluated each combination of axis values; the 17 resulting models¹⁶ are listed in the top part of Table 3. In addition to the 17 exhaustive axis combinations for models with maximum levels in $\{1, 2, 3\}$, we also evaluated (1) 12 additional axis combinations that have higher maximum levels and 8 parameters or fewer: **ai-QCH4** and **ai-QLk4**; **ah-QCH** and **ah-QLk** variations with maximum levels in $\{4, 5, 6, 7\}$; and (2) **ah-QCH** and **ah-QLk** variations that

¹⁴This is especially important when extending the model to higher levels, since there is no general way to represent level-3 agents’ reasoning about level-1 agents simply by varying the precision of the level-3’s predictions about the population.

¹⁵This is in the same spirit as the simplifying assumption made in cognitive hierarchy models that agents have accurate beliefs about the proportions of lower-level agents.

¹⁶When the maximum level is 1, all combinations of the other axes yield identical predictions. Therefore there are only 17 models instead of $3(2^3) = 24$.

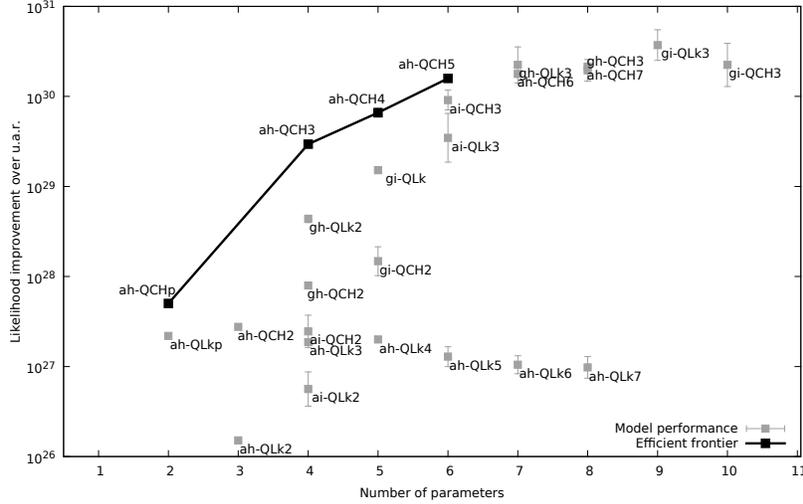


Figure 7: Model simplicity vs. prediction performance. QLk1 is omitted because its far worse performance ($\sim 10^{18}$) distorts the figure’s scale.

assume a Poisson distribution over the levels rather than using an explicit tabular distribution.¹⁷ These additional models are listed in the bottom part of Table 3.

8.1. Simplicity Versus Predictive Performance

We evaluated the predictive performance of each model on the COMBO9 dataset using 10-fold cross-validation repeated 10 times, as in Section 5. The results are given in the last column of Table 3 and plotted in Figure 7.

All else being equal, a model with higher performance is more desirable, as is a model with fewer parameters. We can plot an “efficient frontier” of those models that achieved the (statistically significantly) best performance for a given number of parameters or fewer; see Figure 7. The original QLk model (gi-QLk2) is *not* efficient in this sense; it is dominated by ah-QCH3, which has significantly better predictive performance and fewer parameters (because it restricts agents to homogeneous precisions and accurate beliefs). Our analysis thus argues that the flexibility added by inhomogeneous precisions and general precision beliefs is less important than the number of levels and the choice of population belief.

¹⁷The ah-QCHp model is equivalent to the CH-QRE model of Camerer et al. (2011).

Table 3: Model variations with prediction performance on the COMBO9 dataset. The models with max level of * used a Poisson distribution. Models are named according to precision beliefs, precision homogeneity, population beliefs, and type of level distribution. E.g., **ah-QCH3** is the model with accurate precision beliefs, homogeneous precisions, cognitive hierarchy population beliefs, and a discrete distribution over levels 0–3.

Name	Max Level	Population Beliefs	Precisions	Precision Beliefs	Parameters	Log likelihood vs. u.a.r.
QLk1	1	n/a	n/a	n/a	2	18.37 ± 0.12
gi-QLk2	2	Lk	inhomo.	general	5	29.18 ± 0.03
ai-QLk2	2	Lk	inhomo.	accurate	4	26.75 ± 0.19
gh-QLk2	2	Lk	homo.	general	4	28.64 ± 0.04
ah-QLk2	2	Lk	homo.	accurate	3	26.18 ± 0.03
gi-QCH2	2	CH	inhomo.	general	5	28.17 ± 0.16
ai-QCH2	2	CH	inhomo.	accurate	4	27.39 ± 0.18
gh-QCH2	2	CH	homo.	general	4	27.90 ± 0.03
ah-QCH2	2	CH	homo.	accurate	3	27.44 ± 0.02
gi-QLk3	3	Lk	inhomo.	general	9	30.57 ± 0.17
ai-QLk3	3	Lk	inhomo.	accurate	6	29.54 ± 0.27
gh-QLk3	3	Lk	homo.	general	7	30.35 ± 0.20
ah-QLk3	3	Lk	homo.	accurate	4	27.27 ± 0.03
gi-QCH3	3	CH	inhomo.	general	10	30.35 ± 0.24
ai-QCH3	3	CH	inhomo.	accurate	6	29.96 ± 0.11
gh-QCH3	3	CH	homo.	general	8	30.29 ± 0.12
ah-QCH3	3	CH	homo.	accurate	4	29.47 ± 0.02
ai-QLk4	4	Lk	inhomo.	accurate	8	30.05 ± 0.26
ah-QLk4	4	Lk	homo.	accurate	5	27.30 ± 0.03
ah-QLk5	5	Lk	homo.	accurate	6	27.11 ± 0.11
ah-QLk6	6	Lk	homo.	accurate	7	27.02 ± 0.10
ah-QLk7	7	Lk	homo.	accurate	8	26.99 ± 0.12
ah-QLkp	*	Lk	homo.	accurate	2	27.34 ± 0.02
ai-QCH4	4	CH	inhomo.	accurate	8	29.86 ± 0.20
ah-QCH4	4	CH	homo.	accurate	5	29.82 ± 0.05
ah-QCH5	5	CH	homo.	accurate	6	30.20 ± 0.04
ah-QCH6	6	CH	homo.	accurate	7	30.25 ± 0.03
ah-QCH7	7	CH	homo.	accurate	8	30.33 ± 0.02
ah-QCHp	*	CH	homo.	accurate	2	27.70 ± 0.02

Conversely, the poor performance of the Poisson variants relative to **ah-QCH3** indicates that flexibility in describing the level distribution is more important than the total number of levels modeled.

There is a striking pattern in the models along the efficient frontier: this set consists *exclusively* of models with accurate precision beliefs, homogeneous precisions, and cognitive hierarchy beliefs.¹⁸ This suggests that the most parsimonious way to model human behavior in normal-form games is to use a model

¹⁸One might be interested in a weaker definition of the efficient frontier, saying that a model is efficient if it achieves significantly better performance than all *efficient* models with fewer parameters, rather than *all* models with fewer parameters. In this case the efficient frontier consists of all models previously identified as efficient plus **ah-QCH7** and **gi-QLk3**. Our original definition rejects **gi-QLk3** because it did not predict significantly better than **gh-QLk3**, which in turn did not predict significantly better than **ah-QCH5**.

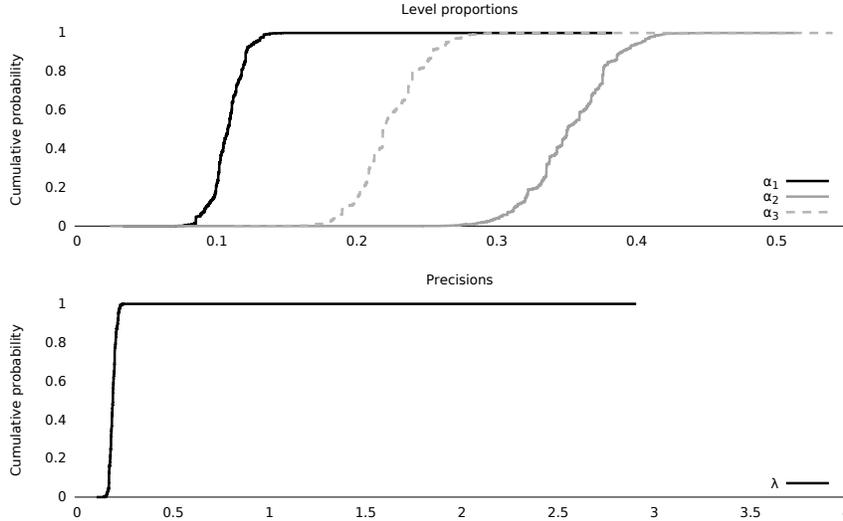


Figure 8: Marginal cumulative posterior distributions for the level proportion parameters ($\alpha_1, \alpha_2, \alpha_3$; top panel) and precision parameter (λ ; bottom panel) of the **ah-QCH3** model on the combined dataset.

of this form, with the tradeoff between simplicity (i.e., number of parameters) and predictive power determined solely by the number of levels modeled. For the **COMB09** dataset, adding additional levels yielded small increases in predictive power until level 5, after which it yielded no further, statistically significant improvements. Thus, Figure 7 includes **ah-QCH4** and **ah-QCH5** as parts of the efficient frontier.

8.2. Parameter Analysis for **ah-QCH3**

We are now in a position to answer some of the questions from Section 7.2 by examining marginal posterior distributions from a member of our new family of “efficient” models (see Figure 8). We first note that, in contrast to **QLk**’s multimodal, jagged parameter CDFs, the parameter CDFs for **ah-QCH3** are smooth and (nearly) unimodal. This suggests that **ah-QCH3** is a much more robust model; its prediction quality is less likely to change drastically as a result of small changes in parameter values.

Second, the posterior distribution for the precision parameter λ is concentrated around 0.20, which is very close to the **QLk** model’s estimate for λ_2 . This

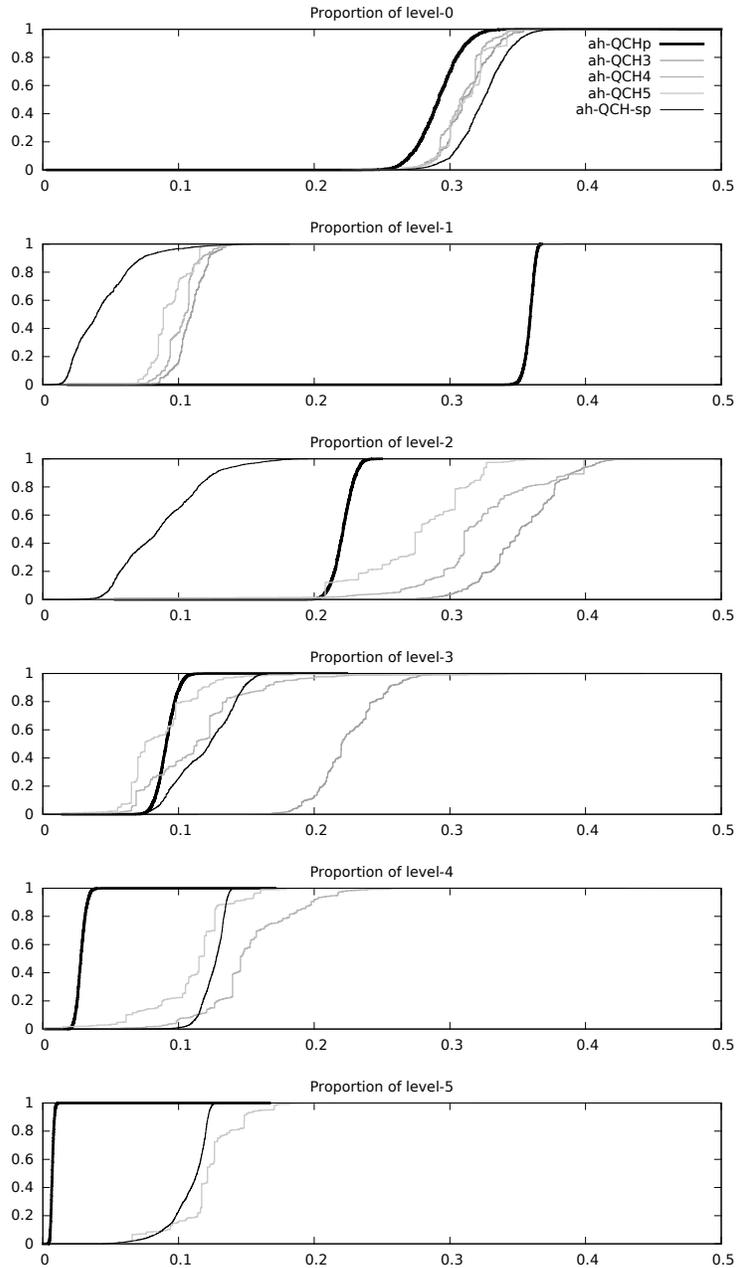


Figure 9: Marginal cumulative posterior distributions of levels of reasoning for efficient frontier models.

suggests that QLk’s much lower estimate for $\lambda_{1(2)}$ may have been the closest that the model could get to having the level-2 agents best respond to a mixture of level-0 and level-1 agents (as in cognitive hierarchy). It is unclear whether the order-of-magnitude differences and counterintuitive ordering of λ_1 and λ_2 are similar effects where QLk’s parameters were set in a way that “simulated” the assumptions of a more accurate model. However, one of our counterintuitive findings is confirmed: as with QLk, for the **ah-QCH3** model we predict more level-2 agents than level-1. In fact, the **ah-QCH3** model predicts *even fewer* level-1 agents than QLk. Furthermore, this prediction appears to be robust across models on our efficient frontier, as illustrated in Figure 9. There is broad agreement among all models on the proportion of level-0 agents, and the tabular-distribution models all select bimodal distributions that assign relatively little weight to level-1 agents, and more to higher-level agents (level-2 and higher). This gives us an intuitive explanation for **ah-QCHp**’s poor performance: it models the level distribution as a (unimodal) Poisson. In order to predict a large number of level-0 agents, **ah-QCHp** must also place substantial weight on level-1 agents.

8.3. Spike-Poisson

We reasoned that a Poisson distribution might be better able to fit our data if the proportion of level-0 agents were specified separately. This would allow for bimodal distributions while still offering the advantage of representing higher-level agents without needing a separate parameter for each level. In this section, we evaluate an **ah-QCH** model that uses just such a distribution: a mixture of a deterministic distribution of level-0 agents and a standard Poisson distribution. We refer to this mixture as a “spike-Poisson” distribution. Because we will end up recommending its use, we define the full Spike-Poisson QCH model here.

Definition 6 (Spike-Poisson Quantal Cognitive Hierarchy (QCH) model). Let $\pi_{i,m}^{SP} \in \Pi(A_i)$ be the distribution over actions predicted for an agent i with level m by the Spike-Poisson QCH model. Let

$$f(m) = \begin{cases} \epsilon + (1 - \epsilon)\text{Poisson}(m; \tau) & \text{if } m = 0, \\ (1 - \epsilon)\text{Poisson}(m; \tau) & \text{otherwise.} \end{cases}$$

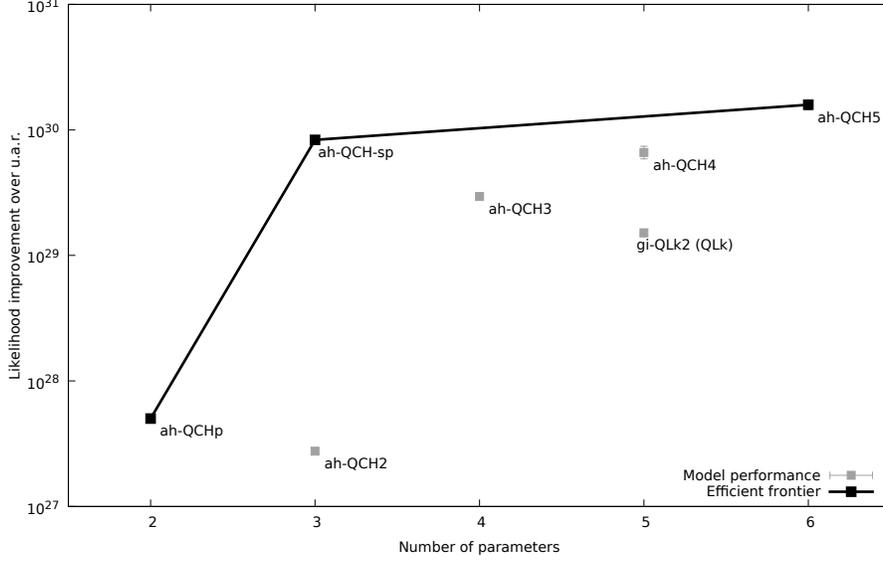


Figure 10: Model simplicity (number of parameters) versus prediction performance on the COMB09 dataset, comparing the ah-QCH models of Section 8.1, QLk, and ah-QCH-sp.

Let $QBR_i^G(s_{-i}; \lambda)$ denote i 's quantal best response in game G to the strategy profile s_{-i} , given precision parameter λ . Let

$$\pi_{i,0:m}^{SP} = \sum_{\ell=0}^m f(\ell) \frac{\pi_{i,\ell}^{SP}}{\sum_{\ell'=0}^m f(\ell')}$$

be the “truncated” distribution over actions predicted for an agent conditional on that agent’s having level $0 \leq \ell \leq m$. Then π^{SP} is defined as

$$\begin{aligned} \pi_{i,0}^{SP}(a_i) &= |A_i|^{-1}, \\ \pi_{i,m}^{SP}(a_i) &= QBR_i^G(\pi_{i,0:m-1}^{SP}). \end{aligned}$$

The overall predicted distribution of actions is a weighted sum of the distributions for each level:

$$\Pr(a_i | \tau, \epsilon, \lambda) = \sum_{\ell=0}^{\infty} f(\ell) \pi_{i,\ell}^{SP}(a_i).$$

The model thus has three parameters: the mean of the Poisson distribution τ , the spike probability ϵ , and the precision λ .

Figure 10 compares the performance of ah-QCH-sp to the ah-QCH models of Section 8.1; for reference, QLk is also included. The three-parameter ah-QCH-sp model outperformed every model except for ah-QCH5. In particular, it outper-

formed both **ah-QCH3** and **ah-QCH4**, despite having fewer parameters than either of them.

The modeling of very high-level agents (i.e., level-5 and higher) seems to determine the performance differences between **ah-QCH-sp** and the other **ah-QCH** models. Hence **ah-QCH-sp**, which includes level-5 agents, outperforms models that do not; but precisely tuning the proportions of levels 5 and below is more important for prediction performance than including levels 6 and above. This is surprising, as agents higher than level-3 are not generally believed to exist in any substantial numbers; for example, Arad and Rubinstein (2012) found no evidence for beliefs of fourth order or higher. One possible explanation for the influence of very high-level responses is that they may be correcting for our overly simple specification of level-0 behavior as uniform randomization. In future work we plan to investigate richer specifications of level-0 behavior.

Overall, we believe that our findings recommend the use of Spike-Poisson QCH for predicting human play in unrepeated normal-form games. It was our best performing model except for **ah-QCH5**, and that model required twice as many parameters to achieve slightly better predictive performance.

8.4. Generalization Performance on Unseen Games

In our performance comparisons thus far, we have used cross-validation at the level of individual datapoints (G_i, a_i) . This means that with very high probability, every game in every testing fold also appeared in the corresponding training set. That is, we never evaluate a model’s predictions on an entirely unseen game. The reader might worry that this undermines our claim that we can use a model fit on one set of games to predict behavior in other, unseen games.

We thus performed an alternative analysis in which we compared the performance of the models of Section 5.1, plus the three “efficient” models from Figure 10, using a modified cross-validation procedure. In this procedure, we divided our combined dataset into folds containing equal numbers of games, with all of the datapoints for a given game belonging to a single fold. Hence, we

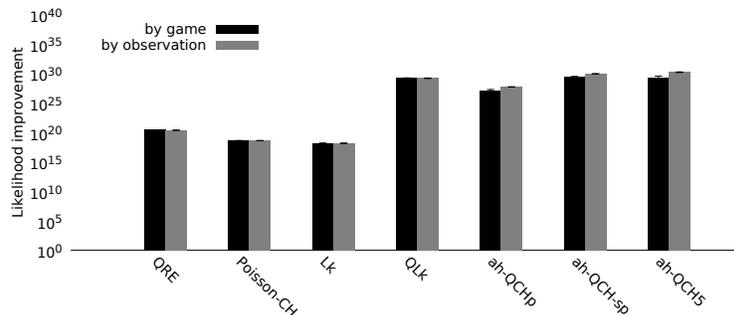


Figure 11: Generalization performance of the QRE, Poisson-CH, Lk, QLk, ah-QCHp, ah-QCH-sp, and ah-QCH5 models on the COMB09 dataset, under two different cross-validation regimes: one in which our data was partitioned into training and test sets at the level of observations, and one in which this partition was performed at the level of games. In the latter condition, models were evaluated entirely using games that were absent from the training data.

evaluated each model entirely using games that were absent from the training set. As before, we report the average of 10 splits into folds to reduce variance in our estimates.

Figure 11 shows the generalization performance of the QRE, Poisson-CH, Lk, QLk, ah-QCHp, ah-QCH-sp, and ah-QCH5 models on the COMB09 dataset under both cross-validation procedures. Overall, performance was virtually identical under the two procedures, suggesting that the models we studied generalize well to unseen games. For the efficient-frontier models we observed small but statistically significant degradations in performance on unseen games; the other models had indistinguishable performance.

9. Related Work

Our work has been motivated by the question, “What model is best for predicting human behavior in general, simultaneous-move games?” Before beginning our study, we conducted an exhaustive literature survey to determine the extent to which this question had already been answered. Specifically, we used Google Scholar to identify all (1698) citations to the papers introducing the QRE, CH, Lk and QLk models (McKelvey and Palfrey, 1995; Camerer et al., 2004; Costa-Gomes et al., 2001; Nagel, 1995; Stahl and Wilson, 1994),

and manually checked every reference. We discarded superficial citations, 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 21 papers (including the four with which we began), which we summarize in Table 4. Overall, we found no paper that compared the predictive performance of all four models. Indeed, there are two senses in which the literature focuses on different issues. First, it appears to be more concerned with *explaining* behavior than with *predicting* it. Thus, comparisons of out-of-sample prediction performance were rare. Here we describe the only exceptions that we found:

- Stahl and Wilson (1995) evaluated prediction performance on 3 games using parameters fit from the other games;
- Morgan and Sefton (2002) and Hahn et al. (2010) evaluated prediction performance using held-out test data;
- Camerer et al. (2004) and Chong et al. (2005) computed likelihoods on each individual game in their datasets after using models fit to the $n - 1$ remaining games;
- Crawford and Iriberri (2007a) 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; and
- Camerer et al. (2011) evaluated the performance of QRE and cognitive hierarchy variants on one experimental treatment using parameters estimated on two separate experimental treatments.

Second, most of the papers compared a single one of the four models (often with variations) to Nash equilibrium. Indeed, only seven of the 22 studies (see the bottom portion of Table 4) compared more than one of the four key models, and none of these considered QLk. Only three of these studies explicitly compared the prediction performance of more than one of the four models (Chong et al., 2005;

Crawford and Iriberry, 2007a; Camerer et al., 2011); the remaining four performed comparisons in terms of training set fit (Camerer et al., 2001; Costa-Gomes and Weizsäcker, 2008; Costa-Gomes et al., 2009; Rogers et al., 2009).

Rogers et al. (2009) proposed a unifying framework that generalizes both Poisson-CH and QRE, and compared the fit of several variations within this framework. Notably, their framework allows for quantal response within a cognitive hierarchy model. Their work is thus similar to our own search over a system of QLk variants, but there are several differences. First, we compared out-of-sample prediction performance, not in-sample fit. Second, Rogers et al. restricted the distributions of types to be grid, uniform, or Poisson distributions, whereas we considered unconstrained discrete distributions. Third, they required different types to have different precisions, while we did not. Finally, we considered level- k beliefs as well as cognitive hierarchy beliefs, whereas they compared only cognitive hierarchy belief models (although their framework in principle allows for both).

One line of work from the computer science literature also meets our criteria of predicting action choices and modeling human behavior (Altman et al., 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 did 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, such machine-learning-based methods could clearly be extended to apply to our setting; investigating their performance would be a worthwhile direction for future work.

10. Conclusions

To our knowledge, ours 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. We found that bounded iterated reasoning and cost-proportional errors are both

Table 4: Existing work. ‘f’ indicates comparison of training sample fit only; ‘t’ indicates statistical tests of training sample performance; ‘p’ indicates evaluation of out-of-sample prediction performance.

Paper	Nash	QLk	Lk	CH	QRE
Stahl and Wilson (1994)	t	t			
McKelvey and Palfrey (1995)	f				f
Stahl and Wilson (1995)	f	p			
Costa-Gomes et al. (1998)	f		f		
Haruvy et al. (1999)		t			
Costa-Gomes et al. (2001)	f		f		
Haruvy et al. (2001)		t			
Morgan and Sefton (2002)	f				p
Weizsäcker (2003)	t				t
Camerer et al. (2004)	f			p	
Costa-Gomes and Crawford (2006)	f		f		
Stahl and Haruvy (2008)		t			
Rey-Biel (2009)	t		t		
Georganas et al. (2010)	f		f		
Hahn et al. (2010)				p	
Camerer et al. (2001)				f	f
Chong et al. (2005)	f			p	p
Crawford and Iriberry (2007a)	p		p		p
Costa-Gomes and Weizsäcker (2008)	f		f		f
Costa-Gomes et al. (2009)	f		f	f	f
Rogers et al. (2009)	f			f	f
Camerer et al. (2011)				p	p

critical ingredients in a predictive model of human game theoretic behavior: the best-performing models we studied (QLk and the QCH family) combine both of these elements.

Bayesian parameter analysis is a valuable technique for investigating the behavior and properties of models, particularly because it is able to make quantitative recommendations for parameter values. We showed how Bayesian parameter analysis can be applied to derive concrete recommendations for the use of an existing model, Poisson-CH, differing substantially from widely cited advice in the literature. We also uncovered anomalies in the parameter settings of the best-performing existing model (QLk), which led us to evaluate systematic variations of its modeling assumptions. In the end, we identified a

new model family (the accurate precision belief, homogeneous-precision QCH models) that allows the modeler to trade off complexity against performance along an efficient frontier of models simply by adjusting a single dimension (the number of levels). Further analysis of this family allowed us to construct a particular three-parameter specification (“Spike-Poisson QCH”, or **ah-QCH-sp**) that outperformed every other model we considered that had fewer than six parameters. We thus recommend the use of this model by researchers wanting to predict human play in (unrepeated) normal-form games.

Our parameter estimates for all of the iterative models included a substantial proportion of level-0 agents. The level-0 model is important for predicting the behavior of all agents in an iterative model; both the level-0 agents themselves, and the higher-level agents whose behavior is grounded in a model of level-0 behavior. In future work, we plan to evaluate the predictive improvements, if any, of richer level-0 models.

References

- Altman, A., Bercovici-Boden, A., and Tennenholtz, M. (2006). Learning in one-shot strategic form games. In *ECML 2006, 17th European Conference on Machine Learning*, pages 6–17.
- Arad, A. and Rubinstein, A. (2012). The 11-20 money request game: A level- k reasoning study. *American Economic Review*, 102(7):3561–3573.
- Becker, T., Carter, M., and Naeve, J. (2005). Experts playing the traveler’s dilemma. Working paper, University of Hohenheim.
- Bishop, C. (2006). *Pattern recognition and machine learning*. Springer.
- Burchardi, K. and Penczynski, S. (2011). Out of your mind: Eliciting individual reasoning in one shot games. Working paper, London School of Economics.
- 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. *Quarterly Journal of Economics*, 119(3):861–898.
- Camerer, C., Nunnari, S., and Palfrey, T. R. (2011). Quantal response and nonequilibrium beliefs explain overbidding in maximum-value auctions. Working paper, California Institute of Technology.
- Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.
- Carvalho, D. and Santos-Pinto, L. (2010). A cognitive hierarchy model of behavior in endogenous timing games. Working paper, Université de Lausanne, Faculté des HEC, DEEP.
- Choi, S. (2012). A cognitive hierarchy model of learning in networks. *Review of Economic Design*, 16(2-3):215–250.
- 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*, pages 203–228.
- Cooper, D. and Van Huyck, J. (2003). Evidence on the equivalence of the strategic and extensive form representation of games. *Journal of Economic Theory*, 110(2):290–308.
- Costa-Gomes, M. and Crawford, V. (2006). Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review*, 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, University of California, San Diego.
- 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. *Journal of the European Economic Association*, 7(2-3):365–376.
- Costa-Gomes, M. A. and Weizsäcker, G. (2008). Stated beliefs and play in normal-form games. *The Review of Economic Studies*, 75(3):729–762.
- Crawford, V. and Iriberri, N. (2007a). Fatal attraction: Salience, naivete, and sophistication in experimental “hide-and-seek” games. *American Economic Review*, 97(5):1731–1750.
- Crawford, V. and Iriberri, N. (2007b). Level- k auctions: Can a nonequilibrium model of strategic thinking explain the winner’s curse and overbidding in private-value auctions? *Econometrica*, 75(6):1721–1770.
- Frey, S. and Goldstone, R. (2011). Going with the group in a competitive game of iterated reasoning. In *2011 Proceedings of the Cognitive Science Society*, pages 1912–1917.
- Gao, X. A. and Pfeffer, A. (2010). Learning game representations from data using rationality constraints. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 185–192.
- Georganas, S., Healy, P. J., and Weber, R. (2010). On the persistence of strategic sophistication. Working paper, University of Bonn.
- Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. CRC press.
- Goeree, J. K. and Holt, C. A. (2001). Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review*, 91(5):1402–1422.
- Goodie, A. S., Doshi, P., and Young, D. L. (2012). Levels of theory-of-mind reasoning in competitive games. *Journal of Behavioral Decision Making*, 25(1):95–108.

- 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.
- Haruvy, E. and Stahl, D. (2007). Equilibrium selection and bounded rationality in symmetric normal-form games. *Journal of Economic Behavior and Organization*, 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.
- Ho, T., Camerer, C., and Weigelt, K. (1998). Iterated dominance and iterated best response in experimental “ p -beauty contests”. *American Economic Review*, 88(4):947–969.
- McKelvey, R., McLennan, A., and Turocy, T. (2007). Gambit: Software tools for game theory, version 0.2007. 01.30.
- McKelvey, R. and Palfrey, T. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38.
- Morgan, J. and Sefton, M. (2002). An experimental investigation of unprofitable games. *Games and Economic Behavior*, 40(1):123–146.
- Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review*, 85(5):1313–1326.
- Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, 11(2):125–139.
- 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. *Games and Economic Behavior*, 65(2):572–585.
- Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Verlag.
- Rogers, B. W., Palfrey, T. R., and Camerer, C. F. (2009). Heterogeneous quantal response equilibrium and cognitive hierarchies. *Journal of Economic Theory*, 144(4):1440–1467.
- Stahl, D. and Haruvy, E. (2008). Level- n bounded rationality and dominated strategies in normal-form games. *Journal of Economic Behavior and Organization*, 66(2):226–232.
- Stahl, D. and Wilson, P. (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior and Organization*, 25(3):309–327.
- Stahl, D. and Wilson, P. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254.
- Turocy, T. (2005). A dynamic homotopy interpretation of the logistic quantal response equilibrium correspondence. *Games and Economic Behavior*, 51(2):243–263.
- 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. *Games and Economic Behavior*, 44(1):145–171.
- Witten, I. H. and Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.
- Wright, J. R. and Leyton-Brown, K. (2010). Beyond equilibrium: Predicting human behavior in normal-form games. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 901–907.

Wright, J. R. and Leyton-Brown, K. (2012). Behavioral game-theoretic models: A Bayesian framework for parameter analysis. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, volume 2, pages 921–928.