Games and Economic Behavior 106 (2017) 16–37

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Predicting human behavior in unrepeated, simultaneous-move games ☆

James R. Wright *, Kevin Leyton-Brown **

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

ARTICLE INFO

Article history:
Received 2 October 2014
Available online 21 September 2017

JEL classification:
C70

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

ABSTRACT

It is commonly assumed that agents will adopt Nash equilibrium strategies; however, experimental studies have demonstrated that this is often a poor description of human players' behavior in unrepeated normal-form games. We analyze five widely studied models of human behavior: Quantal Response Equilibrium, Level-*k*, Cognitive Hierarchy, QLk, and Noisy Introspection. We performed what we believe is the most comprehensive meta-analysis of these models, leveraging ten datasets from the literature recording human play of two-player games. We first evaluated *predictive* performance, asking how well each model fits unseen *test data* using parameters calibrated from separate *training data*. The QLk model (Stahl and Wilson, 1994) consistently achieved the best performance. Using a Bayesian analysis, we found that QLk's estimated parameter values were not consistent with their intended economic interpretations. Finally, we evaluated model variants similar to QLk, identifying one (Camerer et al., 2016) that achieves better predictive performance with fewer parameters.

© 2017 Elsevier Inc. All rights reserved.

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 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 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

☆ 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.
** Corresponding author.
E-mail addresses: jrwright@cs.ubc.ca (J.R. Wright), kevinlb@cs.ubc.ca (K. Leyton-Brown).

https://doi.org/10.1016/j.geb.2017.09.009
0899-8256/© 2017 Elsevier Inc. All rights reserved.

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, five key paradigms have emerged for modeling human decision making in this setting: quantal response equilibrium (QRE; McKelvey and Palfrey, 1995); the noisy introspection model (NI; Goeree and Holt, 2004); the cognitive hierarchy model (CH; Camerer et al., 2004); 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; Weizsäcker, 2003; Rogers et al., 2009), the overwhelming majority of behavioral models of initial play of normal-form games fall broadly into this categorization.

The first contribution of our work is methodological: we demonstrate broadly applicable techniques for comparing and analyzing behavioral models. (See Section 10.1 for our specific methodological recommendations.) We illustrate the use of these techniques via an extensive meta-analysis based on data published in ten different studies, rigorously comparing Lk, QLk, CH, NI, and QRE to each other and to a model based on Nash equilibrium. The findings that result from this meta-analysis both demonstrate the usefulness of the approach and constitute our second contribution. Our first main finding is that QLk is the best performing of these predictive models, both on most individual source datasets and also on a dataset pooling all of the ten datasets. We then analyze and interpret the parameter distributions for several models, including QLk. Based on this analysis, we construct and evaluate a family of variations on QLk. Our second main finding is that a simpler (two-parameter) model achieves better out-of-sample predictive performance than any of the models from the literature that we considered. We recommend the use of this model, dubbed Poisson-QCH, by researchers wanting to predict human play in unrepeated normal-form games.

All of the models we consider 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.[1] 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 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 are behavioral models. Second, two high-level themes that underlie the five 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 consisted entirely of two-player games. 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. Nevertheless, empirically confirming this expectation is an important future direction.

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. 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 three models: a model based on Nash equilibrium, 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 one-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

[1] 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.

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. We defer derivations to appendices. A final appendix investigates the sensitivity of our results to dataset composition, studying how model performance varies with important game properties such as degree of dominance solvability and Nash equilibrium structure.

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 $\Pr(a \mid G, \theta)$. In what follows, we define five prominent behavioral models of human play in unrepeated, simultaneous-move games.[2]

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

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

where λ (the *precision* parameter) indicates how sensitive agents are to utility differences, with $\lambda = 0$ corresponding to uniform randomization and $\lambda \to \infty$ corresponding to best response. When its value is clear from context, we will omit the precision parameter. 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 response 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; i.e., $s_i^* = QBR_i^G(s_{-i}^*; \lambda)$ for all agents i. □

A QRE is guaranteed to exist for any normal-form game and non-negative precision (McKelvey and Palfrey, 1995). However, QRE are 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 $\lambda \to \infty$ 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 are only unique up to affine transformations (Von Neumann and Morgenstern, 1944); hence, equivalent utility functions that have been multiplied by different constants will induce different QREs. The QRE concept nevertheless makes sense if human players are believed to play games differently depending on the magnitudes of the payoffs involved.

[2] We focus here on models of behavior in general one-shot, normal-form games. We omit models of learning in repeated normal-form games such as impulse-balance equilibrium (Selten and Buchta, 1994), payoff-sampling equilibrium (Osborne and Rubinstein, 1998), action-sampling equilibrium (Selten and Chmura, 2008), and experience-weighted attraction (Camerer and Hua Ho, 1999), and models restricted to single game classes, such as cooperative equilibrium (Capraro, 2013). We also omit variants and generalizations of the models we study, such as those introduced by Rogers et al. (2009), Weizsäcker (2003), and Cabrera et al. (2007); however, see Section 8, where we systematically explored a particular space of variants.

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 $k_i \in \{0, 1, 2, \ldots\}$, corresponding to the number of iterations of reasoning the agent is able to perform. A *level-0 agent* plays randomly, choosing uniformly at random from his possible actions. A *level-k agent*, for $k \geq 1$, best responds to the strategy played by level-$(k-1)$ agents. If a level-k agent has more than one best response, he mixes uniformly over them.

We consider a particular level-k model, dubbed Lk, which assumes that all agents belong to levels 0, 1, and 2.[3] 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 let $BR_i^G(s_{-i})$ 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

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

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

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

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

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 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 to $(m-1)$, rather than only to level-$(m-1)$ agents. More formally, every agent has an associated level $m \in \{0, 1, 2, \ldots\}$. 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 $f(j \mid j < m)$.

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

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

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

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

[3] We here model only level-k agents, unlike Costa-Gomes et al. (2001) who also modeled other decision rules. Like Costa-Gomes et al. (2001), we restrict agents' levels to be no greater than 2; however, see Section 8, in which we extend this level-k model to higher levels.

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

The Poisson distribution's mean, τ, is thus this model's single parameter. □

Rogers et al. (2009) note that cognitive hierarchy and QRE often make similar predictions. One possible explanation for this is that cost-proportional errors are adequately captured by cognitive hierarchy (and other iterative models), even though they do not explicitly model this effect. Alternatively, these phenomena could be sufficiently distinct that explicitly modeling both limited iterative strategic thinking and cost-proportional errors yields improved predictions.

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 QLk model (for quantal level-k). In QLk, agents have one of three levels, as in Lk.[4] 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. This 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. The overall predicted distribution of actions is the weighted sum of the distributions for each level,

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

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

2.5. Noisy introspection

Goeree and Holt (2004) propose a model called *noisy introspection* that combines cost-proportional errors and an iterative view of strategic cognition in a different way. Rather than assuming a fixed limit on the number of iterations of strategic thinking, they instead model cognitive bounds by injecting noise into iterated beliefs about others' beliefs and decisions, with the effect that deeper levels of reasoning are assumed to be noisier. They then show that this process of noise injection converges to a unique prediction after a finite number of iterations, which for most games is relatively small.

Goeree and Holt also introduce a concrete version of this model (which we dub NI), in which deeper levels of reasoning are exponentially noisier.

Definition 6 *(NI model).* Define $\pi_{i,k}^{NI,n}$ as

$$\pi_{i,k}^{NI,n} = \begin{cases} QBR_i^G(\pi_{-i,k+1}^{NI,n}; \lambda_0/t^k) & \text{if } k < n, \\ QBR_i^G(p_0; \lambda_0/t^n) & \text{otherwise,} \end{cases}$$

[4] Stahl 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.

where p_0 is an arbitrary mixed profile, $\lambda_0 \geq 0$ is a precision, and $t > 1$ is a "telescoping" parameter that determines how quickly noise increases with depth of reasoning. Then the NI model predicts that each agent will play according to

$$\pi_i^{NI} = \lim_{n \to \infty} \pi_{i,0}^{NI,n}.$$

For a fixed game G, precision λ_0, and telescoping parameter t, this converges to a unique strategy profile regardless of the choice of p_0, since in the limit the precision becomes low enough to bring any profile arbitrarily close to the uniform distribution. □

3. Comparing models

3.1. Prediction framework

How do we determine whether a behavioral model is well supported by experimental data? An experimental dataset $\mathcal{D} = \{(G_i, \{a_{ij} \mid j = 1, \ldots, J_i\}) \mid i = 1, \ldots, I\}$ is a set containing I elements. Each element is a tuple containing a game G_i and a set of J_i pure actions a_{ij}, each played by a human subject in G_i. There is no reason to maintain the pairing of 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 a_i in G_i, which we denote $\Pr(a_i \mid G_i, \theta)$.

A behavioral 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 the model predicted the observations. This takes account of the full predicted distribution; in particular, for any given observed distribution, the prediction that maximizes the likelihood score is the observed distribution itself.[5]

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

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

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[6]:

$$\Pr(a \mid G, \mathcal{D}) = \Pr(a \mid G, \hat{\theta}) = \prod_{i=1}^{I} \prod_{j=1}^{J_i} \Pr(a_{ij} \mid G_i, \theta). \tag{2}$$

3.2. Assessing generalization performance

Each of the models that we consider depends on parameters that are estimated from the data. This presents a problem for evaluating models' performance, since a more flexible model might fit a given dataset better without necessarily predicting unseen data better. Models that perform well by fitting a specific dataset well, but perform poorly at predicting out-of-sample data (i.e., data that was not used for fitting the model's parameters), are said to *overfit* the data.

There are several approaches to avoiding the overfitting problem. One is to compare models' fits to the experimental data, but to apply a penalty to models with larger numbers of parameters. The widely used Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) (e.g., Murphy, 2012) take this approach. However, both criteria are only guaranteed to apply asymptotically in the limit of infinite quantities of data; furthermore, the BIC is only applicable to *nested* models, where one model is a strict generalization of the other. A similar approach is taken by the χ-squared test, which tests the hypothesis that a more-general model's fit is significantly better than that of a restricted model. However, this is difficult to apply to testing multiple models, in addition to again requiring the models to be nested. A third approach to evaluating

[5] Although the likelihood is the quantity that interests us, in practice we operate on the log of the likelihood to avoid numerical precision problems that arise in dealing with exceedingly small quantities. Since log likelihood is a monotonic function of likelihood, a model that has higher likelihood than another model will also have higher log likelihood, and vice versa.

[6] We derive Equation (2) in Appendix A.

predictive performance is to formulate hypotheses based on implications derived directly from a model's definition (see Haile et al., 2008; Hargreaves Heap et al., 2014, for examples of such an approach). This can be a very effective way of evaluating the predictive performance of a single model; however, due to the binary nature of hypothesis testing, it is less appropriate for comparing multiple models.

In this work, we take a fourth approach, which is widespread in machine learning. We estimate parameters on a dataset containing a subset of the data (the *training data*), and then evaluate the resulting model by computing likelihood scores on the observations associated with the remaining, disjoint *test data*. That is, every model's performance is evaluated entirely based on data that were not used for estimating parameters. We partition data at the level of games: data from a given game appears either in the training set or the test set, but not both.[7,8]

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*.[9] Specifically, for each round, we randomly partition the games into 10 parts of approximately equal size. 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 the observations associated with the games belonging to those 9 parts. We then determine the likelihood of the observations in 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 a baseline model based on Nash equilibrium.

4.1. Data

As described in detail in Section 9, we conducted an extensive survey of papers that make use of the five behavioral models we consider.[10] We thereby identified ten large-scale, publicly available sets of human-subject experimental data (Stahl and Wilson, 1994, 1995; Costa-Gomes et al., 1998; Goeree and Holt, 2001; Haruvy et al., 2001; Cooper and Van Huyck, 2003; Haruvy and Stahl, 2007; Costa-Gomes and Weizsäcker, 2008; Stahl and Haruvy, 2008; Rogers et al., 2009). We study all ten[11] of these datasets in this paper. See Table 1 for a summary.

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. 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. The remaining studies consisted exclusively of normal-form games.

All games had two players, so each single play of a game generated two observations. We built one dataset for each study. We also constructed a combined dataset, dubbed ALL10, containing data from all the datasets. The datasets contained very different numbers of observations, ranging from 400 (Stahl and Wilson, 1994) to 2992 (Cooper and Van Huyck, 2003). To ensure that each fold had approximately the same population of subjects, we evaluated ALL10 using *stratified* cross-

[7] This means that observations for a given game will appear in exactly one part of the partition. However, observations from the same subject may appear in multiple parts, when subjects play more than one game.

[8] In an earlier version of this work, we partitioned our dataset at the level of observations. Partitioning at the level of games provides stronger protection against overfitting.

[9] Repeatedly fitting parameters on a bootstrapped subsample and then evaluating performance on the remaining data is another approach to reducing the variance associated with the division into test and training sets. This is a more effective approach for reducing the variance of parameter estimates; however, it introduces bias into performance estimates (Efron and Tibshirani, 1997), which are our primary focus in this work.

[10] One might wonder whether models tended to do better in datasets from studies that explicitly considered them. This turned out not to be the case; a given model's performance in a given individual source dataset had essentially no relationship to whether the source dataset had explicitly studied the model.

[11] We identified an additional dataset (Costa-Gomes and Crawford, 2006) which we do not include due to a computational issue. The games in this dataset had between 200 and 800 actions per player, which made it intractable to compute many solution concepts. As with Nash equilibrium, the main bottleneck in computing behavioral solution concepts is computing expected utilities. Each epoch of training for this dataset requires calculating expected values over up to 640,000 outcomes per game, in contrast to between 9 and approximately 14,000 outcomes per game in the ALL10 dataset. We attempted to overcome this problem by deriving a coarse version of this data by binning similar actions; however, binning in this way resulted in games that were not strategically equivalent to the originals (e.g., when multiple iterations of best response would result in the same binned action in the coarsened games but different unbinned actions in the original games). An open problem for future work is finding a way to address this computational problem by representing the games *compactly* (e.g., Kearns et al., 2001; Koller and Milch, 2001; Jiang et al., 2011), such that expected utility can be computed efficiently over even a very large action space.

Table 1
Names and contents of each dataset. Units are in expected value, in US dollars.

Name	Source	Games	n	Units
SW94	Stahl and Wilson (1994)	10	400	$0.025
SW95	Stahl and Wilson (1995)	12	576	$0.02
CGCB98	Costa-Gomes et al. (1998)	18	1566	$0.022
GH01	Goeree and Holt (2001)	10	500	$0.01
CVH03	Cooper and Van Huyck (2003)	8	2992	$0.10
HSW01	Haruvy et al. (2001)	15	869	$0.02
HS07	Haruvy and Stahl (2007)	20	2940	$0.02
CGW08	Costa-Gomes and Weizsäcker (2008)	14	1792	$0.0107
SH08	Stahl and Haruvy (2008)	18	1288	$0.02
RPC08	Rogers et al. (2009)	17	1210	$0.01
ALL10	Union of above	142	13863	per source

validation: we performed the game partitioning and selection process separately for each of the contained source datasets, thereby ensuring that the number of games from each source dataset was approximately equal in each partition element.

Several studies (Stahl and Wilson, 1994, 1995; Haruvy et al., 2001; Haruvy and Stahl, 2007; Stahl and Haruvy, 2008) paid participants according to a randomized procedure in which experimental subjects played normal-form games for points representing a 1% chance (per game) of winning a cash prize. In Costa-Gomes et al. (1998), each payoff unit was worth 40 cents, but participants were paid based on the outcome of only one randomly-selected game. In the remaining studies (Goeree and Holt, 2001; Cooper and Van Huyck, 2003; Costa-Gomes and Weizsäcker, 2008; Rogers et al., 2009), game payoffs were worth a deterministic number of cents. We summarize the expected value of payoff points in the "Units" column of Table 1. The QRE and QLk models depend on a precision parameter that is not scale invariant. E.g., 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 ALL10 dataset where observations from multiple studies were combined, we normalized the payoff values for each game to be in expected cents.

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 82% of the games in our ALL10 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, the Nash equilibrium model considers our experimental data to be *impossible*, and hence receives a likelihood of zero.

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 by reporting statistics across different equilibria. Specifically, we report the performance achieved by choosing the equilibrium 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, these fits are not able to generalize to new data, and hence cannot be used in practice.) Finally, we also reported the prediction performance on the test data, averaged over all of the Nash equilibria of the game.[12]

4.3. Computational environment

We performed computation using WestGrid (www.westgrid.ca), primarily on the `orcinus` cluster, which has 9600 64-bit Intel Xeon CPU cores. 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).

[12] One might wonder whether the ϵ-equilibrium solution concept (see e.g. Shoham and Leyton-Brown, 2008, Section 3.4.7) solves either of these problems. It does not. First, ϵ-equilibrium can still assign probability 0 to some actions. Second, 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.

5. Model comparisons

In this section we describe the results of our experiments comparing the predictive performance of the five behavioral models from Section 2 and of the Nash-based models of Section 4.2. Fig. 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 ALL10 dataset was approximately 10^{90} times more likely to have been generated by an agent acting according to our Poisson-CH model than choosing actions uniformly at random. 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 across cross-validation partitions; in most cases, these intervals are imperceptibly narrow.

5.1. Comparing behavioral models

Poisson-CH and Lk achieved very similar performance in most datasets. In one way this is an intuitive result, since the models are very similar to each other. On the other hand, it suggests something less obvious, that two differences between the models are not very important in practice: (1) reasoning about just one lower level versus reasoning about the distribution of all lower levels; (2) the distinct error models.

QRE and NI tended to perform well on the same datasets. On all but two datasets (HSW01 and CGW08), the ordering between QRE and the iterative models was the same as between NI and the iterative models. We found this result surprising, since the two models appear quite different. However, the two models do share several key elements in common. First, both models are based around cost-proportional errors, and they both assume that all agents play from the same distribution, unlike the iterative models, which assume that different agents reason to different depths. Further, although NI is not explicitly a fixed-point model, it does assume an unlimited depth of reasoning, like QRE, although it does typically converge after a relatively small number of iterations.

In five datasets, the models based on cost-proportional errors (QRE and NI) predicted human play significantly better than the two models based on bounded iterated reasoning (Lk and Poisson-CH). However, in five other datasets, including ALL10, the situation was reversed, with Lk and Poisson-CH outperforming QRE and NI. In the remaining two datasets, NI outperformed the iterative models, which outperformed QRE. This mixed result is consistent with earlier, less extensive comparisons of QRE with these two models (Chong et al., 2005; Crawford and Iriberri, 2007a; Rogers et al., 2009, see also Section 9), 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 we were right about this hypothesis, we might expect that our remaining model, which incorporates both components, would predict better than models that are based on only one component. This was indeed the case: QLk generally outperformed the single-component models. Overall, QLk was the strongest behavioral model; in a majority of datasets, no model made significantly better predictions. The datasets in which some model other than QLk did make significantly better predictions were CVH03, SW95, CGCB98, and GH01; we discuss the latter in detail below, in Section 5.2.

We typically estimated different parameter values than the papers that introduced the models we studied. One reason[13] this occurred is that our training set contains a only subset of these games. This sensitivity to taking subsets of games indicates that overfitting is indeed a realistic concern.

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., Goeree and Holt, 2001). Nevertheless, for the sake of completeness, we also evaluated the predictive power of Nash equilibrium with error (NEE) on our datasets. Referring again to Fig. 1, we see that NEE's predictions were worse than those of every behavioral model on every dataset except SW95 and CGCB98. NEE's upper bound—using the post-hoc best equilibrium—was significantly worse than QLk's performance on every dataset except SW95, CGCB98, RPC09, 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.[14] Fig. 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

[13] In at least one case, our values are also different due to errors in an original paper's estimation: Stahl and Wilson (1994) estimated level proportions that sum to more than 1.

[14] Of course, GH01 was also constructed so that human play on the other half of its games would be poorly described by Nash equilibrium. However, this is still a difference from the other datasets, in which Nash equilibrium appears to have poorly described an even larger fraction of games.

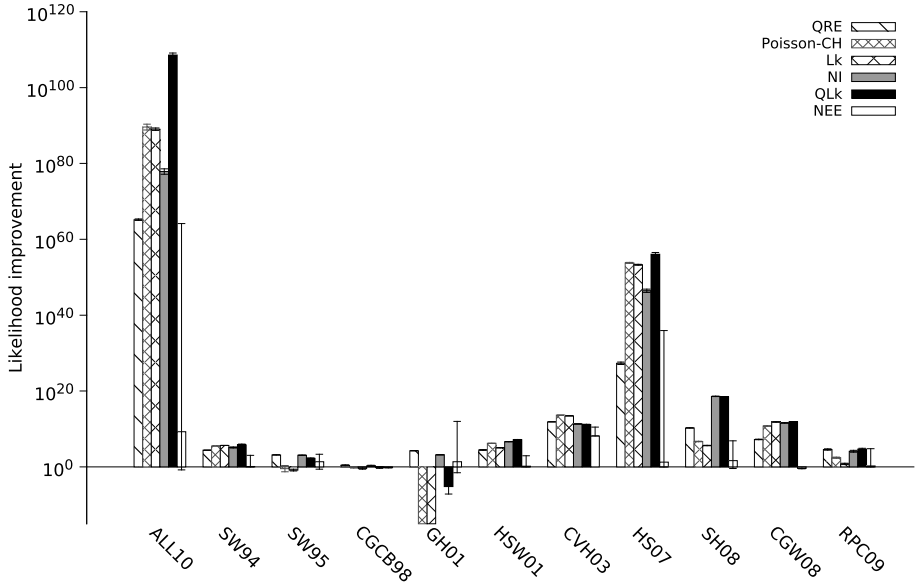

Fig. 1. Average likelihood ratios of model predictions to random predictions, with 95% confidence intervals. Error bars for NEE show upper and lower bounds on performance depending upon equilibrium selection; the main bar for NEE shows the average performance over all equilibria. Note that conclusions should not be drawn about relative differences in likelihood across datasets, as likelihood depends on the dataset's number of samples and the underlying games' numbers of actions. Relative differences in likelihood *are* meaningful within datasets.

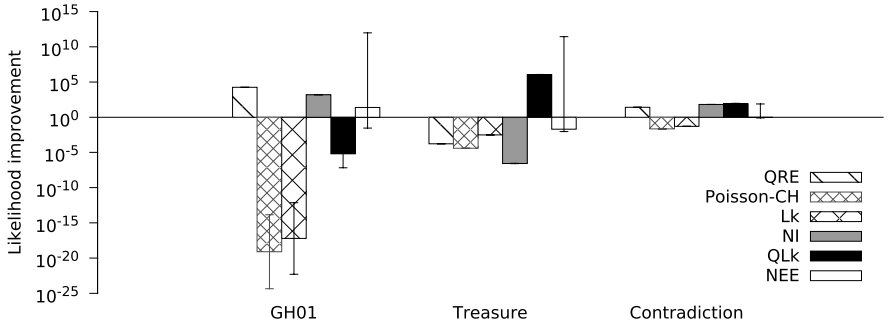

Fig. 2. Average likelihood ratios of model predictions to random predictions, with 95% confidence intervals, on GH01 data separated into "treasure" and "contradiction" treatments. Error bars for NEE show upper and lower bounds on performance depending upon equilibrium selection; the main bar for NEE shows the average performance over all equilibria. Note that relative differences in likelihood are not meaningful across datasets, as 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.

GH01's games have multiple equilibria. This conferred an advantage to our NEE model's upper bound, because it was allowed to pick the equilibrium with best test-set performance on a per-instance basis. Note that although NEE thus had a higher upper bound than QLk on the "treasure" treatment, its average performance was still quite poor.

6. Analyzing model parameters

Making good predictions from behavioral models depends upon obtaining good estimates of model parameters. 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 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. 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.[15]

[15] We can gain local information about 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.

For example, in the previous chapter we noted that our parameter estimates for QLk implied a much larger proportion of level-0 agents than is conventionally expected. We also interpreted the large estimated value of the noise parameter ϵ as indicating that Nash equilibrium fits the data poorly. However, much less can be concluded from such facts if there turn out to be multiple, very different ways of configuring these models to make good predictions.

An alternative is to use Bayesian analysis to estimate the entire posterior distribution over parameter values rather than estimating only a single point. 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. We derive an expression for the posterior distribution in Appendix B. In Section 7 we will apply these methods to study QLk, NEE, and Poisson-CH: the first because it achieved such reliably strong performance; the second because it has an error term with an especially interpretable posterior distribution; and the last because it is the model about which the most explicit parameter recommendation was made in the literature. Camerer et al. (2004) recommended setting Poisson-CH'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 roughly a factor of two lower, on the range $[0.70, 0.76]$. We devote most of our attention to QLk, however, due to its extremely strong performance.

6.1. 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 a 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. For models such as QLk with a multidimensional parameter space, we used *Metropolis–Hastings sampling* to estimate the posterior distribution. The Metropolis–Hastings algorithm is a Markov Chain Monte Carlo (MCMC) algorithm (e.g., Robert and Casella, 2004) that computes a series of values from the support of a distribution. Although each value depends upon the previous value, the values are distributed as if from an independent sample of the distribution after a sufficiently large number of iterations. MCMC algorithms (and related techniques, e.g., annealed importance sampling, Neal, 2001) are useful for estimating multidimensional distributions for which a closed form of the density is unknown. They require only that a value *proportional* to the true density be computable (i.e., an unnormalized density). This is precisely the case with the models that we seek to estimate.

We used a flat prior for all parameters.[16] Although this prior is improper on unbounded parameters such as precision, it results in a correctly normalized posterior distribution[17]; the posterior distribution in this case reduces to the likelihood (e.g., 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 turned out not to matter, as the vast majority of probability mass was concentrated near 0.

6.2. 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 further, unpublished experiments we found little in the way of interaction effects between parameters.

7. Parameter importance analysis

In this section we analyze the posterior distributions of the parameters for three of the models compared in Section 5: Poisson-CH, NEE, and QLk. We then compare our estimates of the relative proportions of level-0 agents to previous work.

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 NEE, we computed the likelihood for each value of $\epsilon \in \{0.01k \mid k \in \mathbb{N}, 0 \leq 0.01k \leq 1\}$. For

[16] For precision parameters, another natural choice might have been to use a flat prior on the log of precision. We chose as we did 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.

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

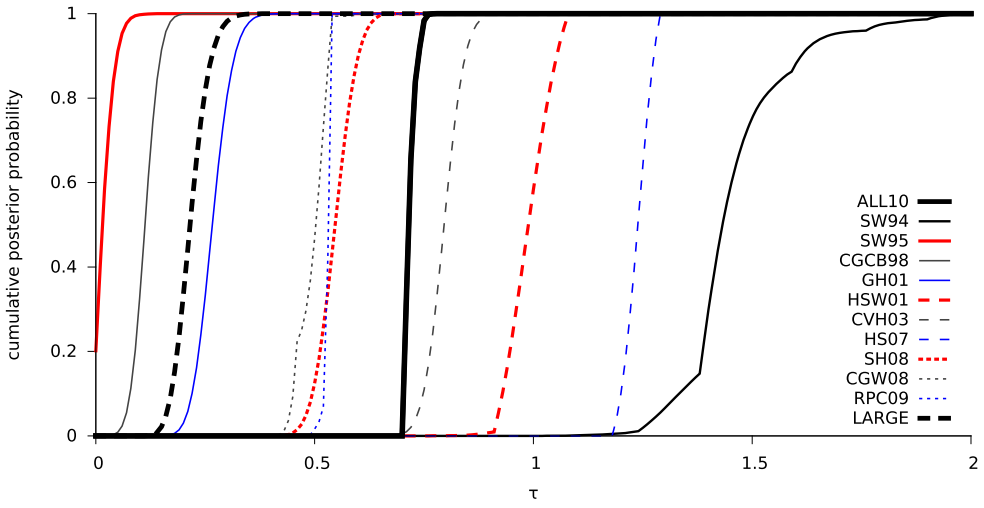

Fig. 3. Cumulative posterior distributions for Poisson-CH's τ parameter. Bold solid trace is the combined dataset; solid black trace is the outlier Stahl and Wilson (1994) source dataset; bold dashed trace is a subset containing all large games (those with more than 5 actions per player).

Lk and QLk, we combined the samples from 4 independent Metropolis-Hastings chains, each of which computed 220,000 samples, discarding the first 20,000 samples as a "burn-in" period to allow the Markov chain to converge. We used the PyMC software package to generate the samples (Patil et al., 2010). Computing the posterior distribution for a single model in this way typically required approximately 200 CPU hours.

7.1. Poisson-CH

In an influential recommendation from the literature, Camerer et al. (2004) suggest[18] 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 Fig. 3). Overall, our analysis strongly contradicts Camerer et al.'s recommendation. On ALL10, the posterior probability of $0.70 \leq \tau \leq 0.76$ is more than 99%. Every other source dataset had a wider 99% *credible interval* (the Bayesian counterpart to confidence intervals) for τ than ALL10, as indicated by the higher slope of ALL10'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, possibly due to the dataset's small size.

Many of the games in our dataset have small action spaces. For example, 108 out of the 142 games in ALL10 have exactly 3 actions per player. One might worry that the estimated average cognitive level in Fig. 3 is artificially low, since it is impossible to distinguish higher numbers of levels than the number of actions available to each player. We check this by performing the same posterior estimation on a subset of the data consisting only of the 4 large games (i.e., those with more than 5 actions available to each player). As Fig. 3 shows, the estimated average cognitive level in these large games was even lower than the overall estimate, with a median of 0.22.

7.2. Nash equilibrium

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 this parameter to have a relatively low value; in contrast, the values of ϵ that maximize NEE's performance were extremely high. In this section we estimate the full posterior distribution for ϵ; see Fig. 4. By doing so we are able to confirm that in both ALL10 and its component source datasets, the posterior distribution for ϵ is very concentrated around very large values of ϵ. 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. This is especially true as the agents within a Nash equilibrium do not take others' noisiness into account, which makes it difficult to interpret ϵ as a measure of level-0 play rather than of model misspecification.

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

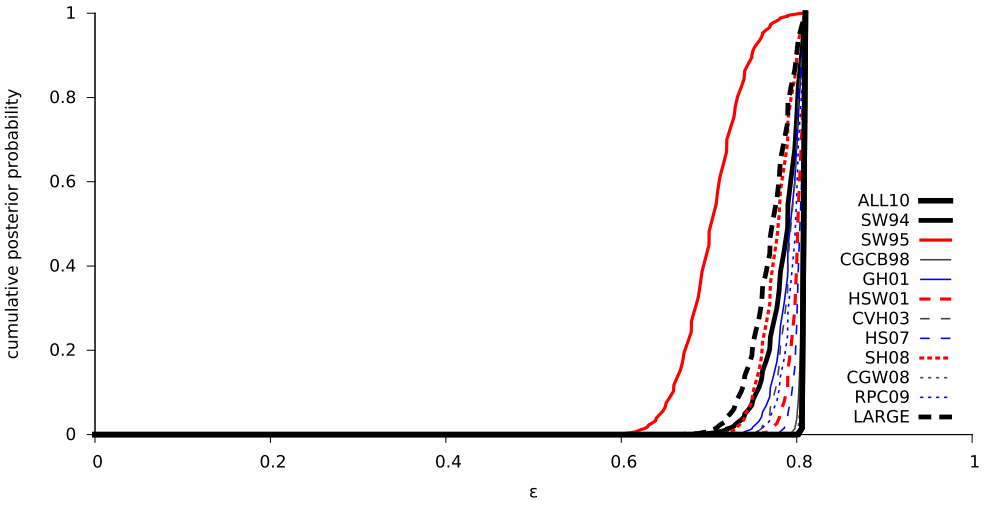

Fig. 4. Cumulative posterior distributions for NEE's ϵ parameter. Bold solid trace is the combined dataset; bold dashed trace is a subset containing all large games (those with more than 5 actions per player).

7.3. QLk

Fig. 5 gives the marginal cumulative posterior distributions for QLk's level proportion distributions broken down by source dataset. That is, we computed the five-dimensional posterior distribution, and then extracted from it the three marginal distributions shown here.[19] As with Poisson-CH, posterior level distributions varied across datasets.[20]

We observe a surprisingly high posterior frequency of level-0 agents. The posterior medians for the proportion of level-0, level-1, and level-2 agents in the ALL10 dataset are 0.32, 0.42, and 0.26, respectively. See Section 7.4 for a further discussion of our level-0 estimates.

Overall, we observed rather small quantal response precisions. In the ALL10 dataset, the posterior median precisions for level-1 agents, level-2 agents, and the belief of level-2 agents about level-1 agents were 0.16, 0.56, and 0.05 respectively. The belief of the level-2 agents that the level-1 agents have a much smaller precision than their actual precision was particularly strongly identified. That is, the ALL10 dataset assigned the highest posterior probability to parameter settings in which the level-2 agents ascribe a smaller than accurate quantal response precision to the level-1 agents. QLk may get this right: e.g., two-level strategic reasoning might cause a high cognitive load, making agents more likely to make mistakes in their predictions of others' behavior. Alternately, we might worry that QLk fails to capture some crucial aspect of experimental subjects' strategic reasoning. For example, the low value of $\lambda_{1(2)}$ might reflect level-2 agents' reasoning about all lower levels rather than just one level below themselves: ascribing a low precision to level-1 agents approximates a mixture of level-1 agents and uniformly randomizing level-0 agents. That is, the low value of $\lambda_{1(2)}$ may be a way of simulating a cognitive hierarchy style of reasoning within a level-k framework. In the next section, we will explore this possibility as part of an evaluation of systematic variations of QLk's modeling assumptions.

7.4. Level-0

Earlier studies found support for widely varying proportions of level-0 agents. Stahl and Wilson (1994) estimated that 0% of the population was level-0[21]; 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 (2014) estimated 37% by fitting a level-k model, and between 20–42% by eliciting subject strategies.

The posterior median for the proportion of level-0 agents in the ALL10 dataset according to the QLk model is 32%, with a 95% credible interval of [29%, 35%]. This is toward the high end of the range of previous estimates. However, note that our estimate for QLk is very similar to the fitted estimate of Burchardi and Penczynski (2014), and comfortably within the range that they estimated by directly evaluating subjects' elicited strategies in a single game. According to the Lk model, the posterior median for the proportion of level-0 agents in ALL10 is 18%. However, the Lk model suffers from an identifiability

[19] We omit marginal distributions for the precision parameters λ_1, λ_2, and $\lambda_{1(2)}$ for space reasons. They follow the same broad pattern as the level proportion distributions: the parameters have relatively diverse posterior distributions and degrees of identification in the individual datasets, but are very sharply identified in the combined ALL10 dataset.

[20] To confirm that these results were not simply an artifact of a difficult-to-sample posterior distribution, we simulated data from ALL10 from a QLk model with known parameters, and then sampled from the posterior distribution of this synthesized dataset. For all 5 parameters, the true parameter value was contained within the 95% central credible interval a minimum of 93 times out of 100 repetitions, indicating that the sampler was well calibrated.

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

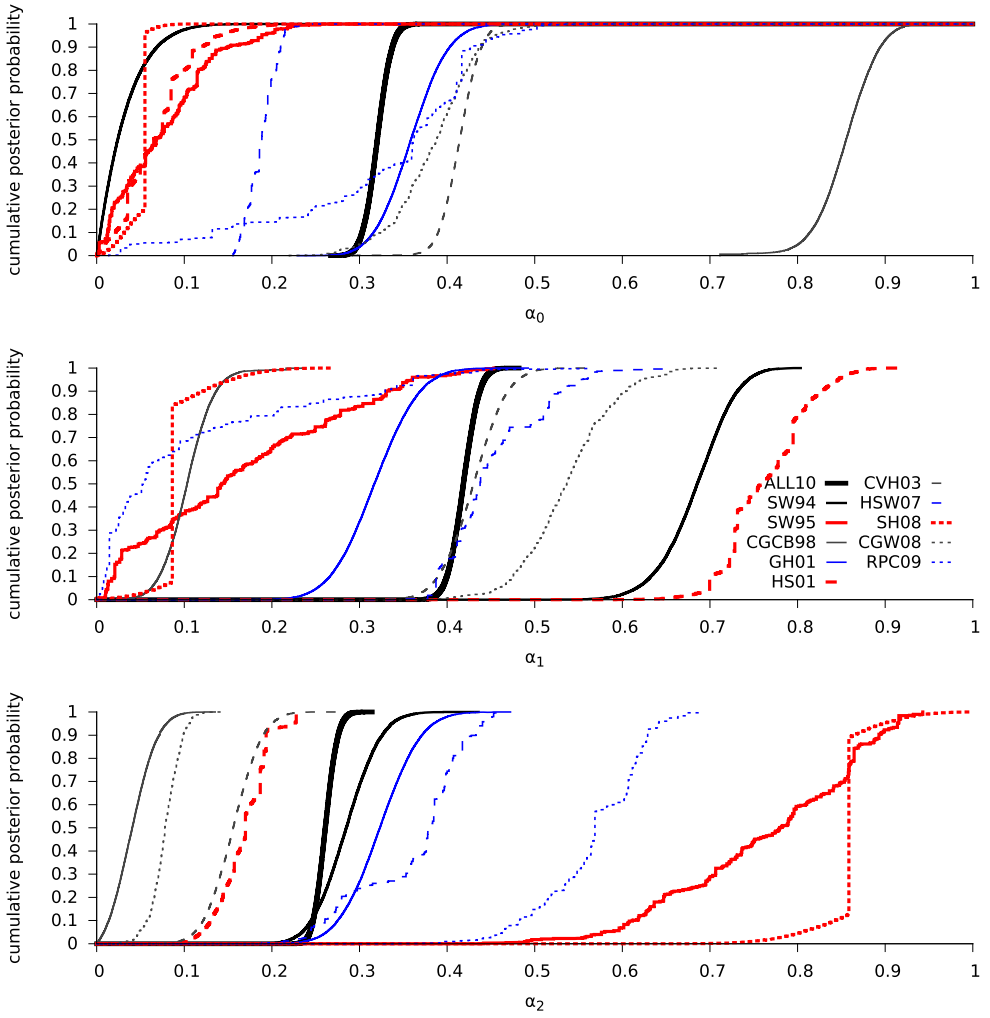

Fig. 5. Marginal cumulative posterior distribution functions for the level proportion parameters $(\alpha_0, \alpha_1, \alpha_2)$ of the QLk model.

problem, in that there is no way to distinguish uniform noise that is introduced by the uniform error structure from uniform noise introduced by level-0 agents. This results in a very wide 95% credible interval of [1%, 42%].

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—that nonstrategic behavior occurs at a substantial frequency—must be taken seriously.

8. Model variations

QLk makes various modeling assumptions that may seem arbitrary. For example, is it the right choice to model exactly two cognitive levels? And, is it really necessary to model the fact that agents at one level might be incorrect about the precision of the level below them? We now investigate these and other such questions, considering a family of models that systematically vary the assumptions underlying QLk. In the end, we identify a simpler model that dominated QLk on our data.

More specifically, we considered four different axes along with the QLk model could be modified. 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

Table 2
Model variations with prediction performance on the ALL10 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	Precision beliefs	Precisions	Parameters	Log likelihood vs. u.a.r.
QLk1	1	n/a	n/a	n/a	2	87.37 ± 1.04
gi-QLk2	2	Lk	general	inhomo.	5	108.66 ± 0.56
ai-QLk2	2	Lk	accurate	inhomo.	4	103.33 ± 1.75
gh-QLk2	2	Lk	general	homo.	4	107.96 ± 0.46
ah-QLk2	2	Lk	accurate	homo.	3	104.84 ± 0.58
gi-QCH2	2	CH	general	inhomo.	5	107.78 ± 0.88
ai-QCH2	2	CH	accurate	inhomo.	4	106.76 ± 0.92
gh-QCH2	2	CH	general	homo.	4	109.43 ± 0.58
ah-QCH2	2	CH	accurate	homo.	3	106.67 ± 0.41
gi-QLk3	3	Lk	general	inhomo.	9	113.17 ± 1.46
ai-QLk3	3	Lk	accurate	inhomo.	6	109.62 ± 1.21
gh-QLk3	3	Lk	general	homo.	7	113.48 ± 1.46
ah-QLk3	3	Lk	accurate	homo.	4	107.12 ± 0.46
gi-QCH3	3	CH	general	inhomo.	10	113.01 ± 0.93
ai-QCH3	3	CH	accurate	inhomo.	6	111.34 ± 0.59
gh-QCH3	3	CH	general	homo.	8	113.08 ± 0.83
ah-QCH3	3	CH	accurate	homo.	4	110.42 ± 0.46
ai-QLk4	4	Lk	accurate	inhomo.	8	110.30 ± 0.93
ah-QLk4	4	Lk	accurate	homo.	5	106.63 ± 0.71
ah-QLk5	5	Lk	accurate	homo.	6	107.18 ± 0.57
ah-QLk6	6	Lk	accurate	homo.	7	106.57 ± 0.68
ah-QLk7	7	Lk	accurate	homo.	8	106.50 ± 0.69
ah-QLkp	$*$	Lk	accurate	homo.	2	106.89 ± 0.28
ai-QCH4	4	CH	accurate	inhomo.	8	111.54 ± 0.62
ah-QCH4	4	CH	accurate	homo.	5	110.88 ± 0.33
ah-QCH5	5	CH	accurate	homo.	6	111.22 ± 0.39
ah-QCH6	6	CH	accurate	homo.	7	111.26 ± 0.44
ah-QCH7	7	CH	accurate	homo.	8	111.42 ± 0.41
ah-QCHp	$*$	CH	accurate	homo.	2	110.48 ± 0.25

make the simplifying assumption that all agents have *accurate precision beliefs* about lower-level agents.[22] 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 models[23] are listed in the top part of Table 2. 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 assume a Poisson distribution over the levels rather than using an explicit tabular distribution.[24] These additional models are listed in the bottom part of Table 2.

8.1. Simplicity versus predictive performance

We evaluated the predictive performance of each model on the ALL10 dataset using 10-fold cross-validation repeated 10 times, as in Section 5. The results are given in the last column of Table 2 and plotted in Fig. 6.

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 best performance for a given number of parameters or fewer; see Fig. 6. The original QLk model (`gi-QLk2`) is *not* efficient in this sense; it is dominated by, e.g., `ah-QCH3`, which has both significantly better predictive performance and fewer parameters (because it restricts agents to homogeneous precisions and accurate beliefs).

There is a striking pattern among the efficient models with 6 parameters or fewer: every such model has accurate precision beliefs, cognitive hierarchy population beliefs, and, with the exception of `ai-QCH3`, homogeneous precisions. Furthermore, `ai-QCH3`'s performance was not significantly better than that of `ah-QCH5`, which did have homogeneous

[22] This 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.

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

[24] The `ah-QCHp` model is identical to the CH-QRE model of Camerer et al. (2016).

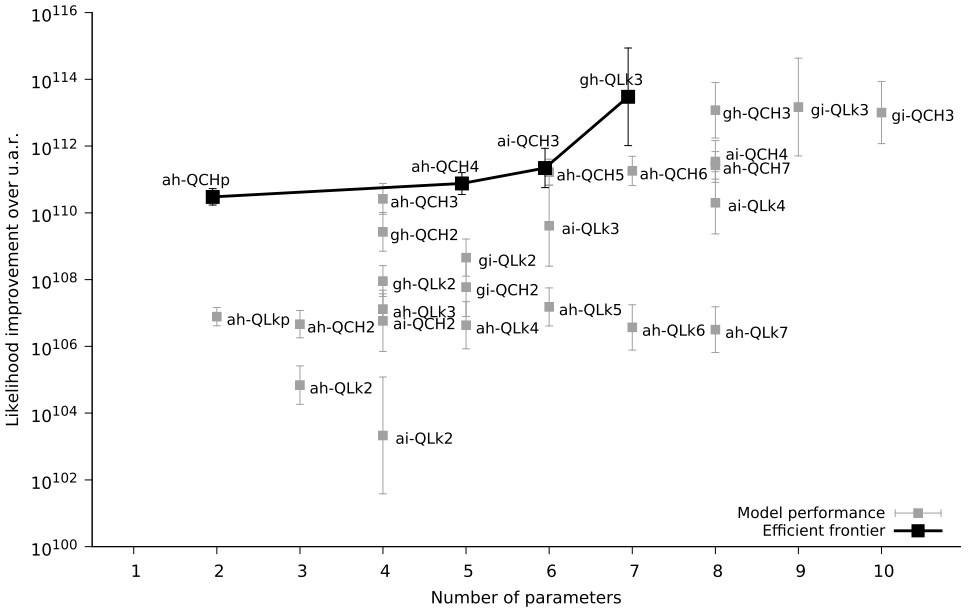

Fig. 6. Model simplicity vs. prediction performance on the ALL10 dataset. QLk1 is omitted because its far worse performance ($\sim 10^{87}$) distorts the figure's scale.

precisions. This suggests that the most parsimonious way to model human behavior in normal-form games is to use a model of this form.

Adding flexibility by modeling general beliefs about precisions did improve performance; the four best-performing models all incorporated general precision beliefs. However, these models also had much larger variance in their prediction performance on the test set. This may indicate that the models are overly flexible, and hence prone to overfitting.

8.2. Parameter analysis of *ah-QCH* models

In this section we examine the marginal posterior distributions of two models from the accurate, homogeneous QCH family (see Fig. 7). We computed the posterior distribution of the models' parameters using the procedure described in Sections 6.1 and 7. The posterior distribution for the precision parameter λ was concentrated around 0.20, somewhat greater than the QLk model's estimate for λ_1. This suggests that QLk's much lower estimate for $\lambda_{1(2)}$ may indeed 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).

Our robust finding in Sections 7.4 and 7.3 of a large proportion of level-0 agents was confirmed by these models as well. Indeed, the number of level-0 agents was nearly the only point of close agreement between all three models with respect to the distribution of levels.

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 (1805) citations to the papers introducing the QRE, CH, Lk, NI, and QLk models (McKelvey and Palfrey, 1995; Camerer et al., 2004; Costa-Gomes et al., 2001; Nagel, 1995; Goeree and Holt, 2004; 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 24 papers, including the six with which we began, which we summarize in Table 3. Overall, we found no paper that compared the predictive performance of all six 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;

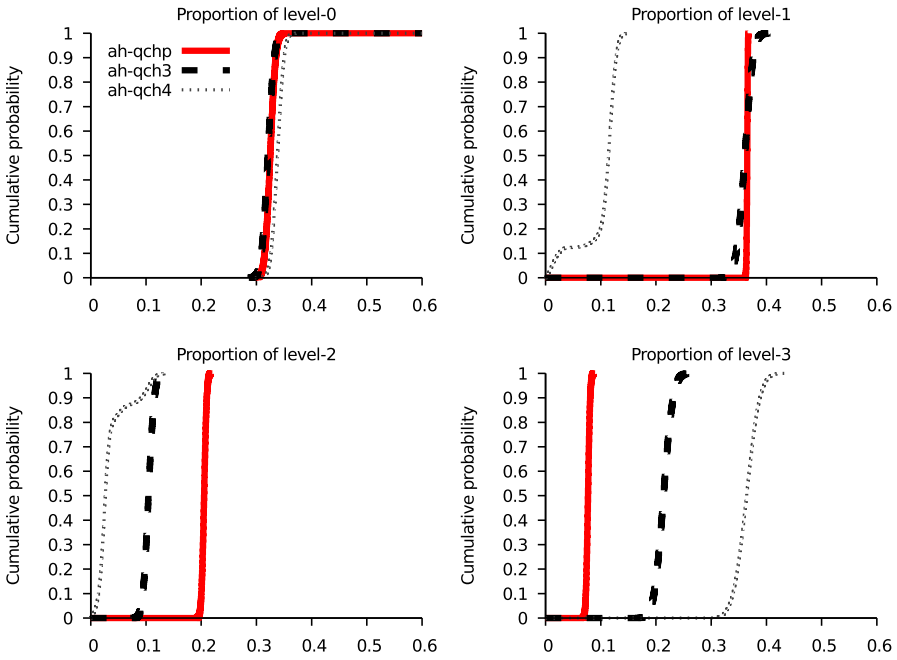

Fig. 7. Marginal cumulative posterior distributions for the level proportion parameters $(\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ of the `ah-QCHp`, `ah-QCH3`, and `ah-QCH4` models on All10. Solid lines are `ah-QCHp`; dashed lines are `ah-QCH3`; dotted lines are `ah-QCH4`. All α values are defined implicitly by the τ parameter for `ah-QCHp`. For the other models, α_0 is defined implicitly by $\alpha_1, \alpha_2, \alpha_3$, and (for `ah-QCH4`) α_4.

- 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. (2016) 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 five models (often with variations) to Nash equilibrium. Indeed, only nine of the 24 studies (see the bottom portion of Table 3) compared more than one of the six key models, and none of these considered QLk. Only three of these studies explicitly compared the prediction performance of more than one of the six models (Chong et al., 2005; Crawford and Iriberri, 2007a; Camerer et al., 2016); the remaining six performed comparisons in terms of training set fit (Camerer et al., 2001; Goeree and Holt, 2004; Costa-Gomes and Weizsäcker, 2008; Costa-Gomes et al., 2009; Rogers et al., 2009; Breitmoser, 2012).

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 in Section 8, 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 over levels. 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 considered only cognitive hierarchy belief models.

One line of work in computer science 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.

10. Conclusions

To our knowledge, ours is the first study to address the question of which existing behavioral model—QRE, level-k, cognitive hierarchy, noisy introspection, or quantal level-k behavioral models—is best suited to predicting unseen human initial 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 valuable ingredients in a predictive model of human game theoretic behavior: the best-performing model that we studied (QLk) combines both of these elements. We believe that iterative reasoning describes an actual cognitive process. The situation is less clear with cost-proportional er-

Table 3
Existing work in model comparison. '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	NI	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. (2015)	f		f			
Hahn et al. (2010)				p		
Camerer et al. (2001)				f		f
Goeree and Holt (2004)	f				f	f
Chong et al. (2005)	f			p		p
Crawford and Iriberri (2007a)	p		p			p
Costa-Gomes and Weizsäcker (2008)	f		f		f	f
Costa-Gomes et al. (2009)	f		f	f	f	f
Rogers et al. (2009)	f			f		f
Camerer et al. (2016)				p		p
Breitmoser (2012)			t	t	t	t

rors: they may likewise describe human reasoning, or they may simply be a closer approximation to human behavior than the usual uniform error specification.

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 Poisson-CH, differing substantially from widely cited advice in the literature.

QLk (`gi-qlk2`) provides substantial flexibility in specifying the beliefs and precisions of different types of agents. We found that this flexibility tends to hurt generalization performance more than it helps. In a systematic search of model variations, we identified a new model family (the accurate precision belief, homogeneous-precision QCH models) that contained the efficient (or nearly-efficient) model for every number of parameters smaller than 7. Based on further analysis of this model family, we identified a model, Poisson-QCH, that offers excellent generalization performance with only two parameters.

10.1. Recommendations

Methodology In this work we have focused exclusively on prediction performance. One might wonder whether there is any practical difference between in-sample fit and out-of-sample prediction performance. It turns out that the ranking of a model's performance within a dataset was identical in the test and training sets only 45% of the time, despite the low dimensionality of the models that we considered. The average difference between a model's rank by test performance and its rank by training performance was 1.5. The `ai-qlk4` model was an especially notable example, having the 5th-highest training performance but only the 14th-highest test performance.

We thus conclude that there is no substitute for evaluating a model on held-out test data. We recommend the use of 10-fold cross-validation, repeated 10 times with a different random partition over games on each repetition, as described in Section 3.2. However, we recognize that this process is computationally intensive, as it requires each model to be fit 100 times. If computation time is a major constraint, we recommend a single round of 10-fold cross-validation, or even a single round of 4-fold cross-validation; this still gives an unbiased estimate of prediction performance, albeit without error bars.

The log-likelihood performance measure has some problematic features: it is not comparable between datasets, and its units do not have an especially natural interpretation. Nevertheless, it is the most appropriate performance measure for predictive behavioral models of which we are aware, especially when normalized against a baseline such as the performance of uniform predictions.

Models Section 8 analyzes an "efficient frontier" of models, each of which represent a different tradeoff between performance and parsimony (and hence robustness). The Poisson-QCH model (`ah-QCHp`) is attractive for being low-variance

and reasonably performant, whereas gh-QLk3 has the highest expected performance but also the highest variance of any model, and a more difficult-to-interpret parameter structure.

We recommend the use of the Poisson-QCH model for the prediction of human strategic behavior in unrepeated, simultaneous-move games.[25] The median posterior parameters for the ALL10 dataset were $\lambda = 0.20, \tau = 1.12$.[26] These settings may be a good starting point for applications, although we note that application-specific fits are always preferable due to behavioral variation across subject populations.

10.2. Further directions

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 ongoing work, we are investigating richer specifications of level-0 behavior, which allow for significant performance improvements (Wright and Leyton-Brown, 2014).

Our approach of fitting the parameters of an iterative model in one set of games and then using these parameters to make predictions in distinct games implicitly assumes that the distribution of beliefs in the population is constant across different games. In several studies, experimental subjects do exhibit surprising stability (Stahl and Wilson, 1994, 1995; Costa-Gomes et al., 2001; Polonio et al., 2015) or convergence (Breitmoser et al., 2014) in their apparent levels of reasoning. However, it also seems reasonable to suppose that players' depths of reasoning would be influenced by the structure of the game. In ongoing work, we are investigating ways to model such endogenous reasoning steps.

Acknowledgments

This work was funded in part by the Natural Sciences and Engineering Research Council of Canada. It was completed in part while the authors were visiting the Simons Institute for the Theory of Computing. We thank several anonymous reviewers and editors for many helpful comments that have significantly improved the paper.

Appendix A. Likelihood derivation

The likelihood of a single datapoint $d_{ij} = (G_i, a_{ij})$ is

$$\Pr(d_{ij} \mid \theta) = \Pr(G_i, a_{ij} \mid \theta).$$

By the chain rule of probabilities, this[27] is equivalent to

$$\Pr(d_{ij} \mid \theta) = \Pr(a_{ij} \mid G_i, \theta) \Pr(G_i \mid \theta),$$

and by independence of G and θ we have

$$\Pr(d_{ij} \mid \theta) = \Pr(a_{ij} \mid G_i, \theta) \Pr(G_i). \tag{A.1}$$

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

$$\Pr(\mathcal{D} \mid \theta) = \prod_{i=1}^{I} \prod_{j=1}^{J_i} \Pr(a_{ij} \mid G_i, \theta) \Pr(G_i). \tag{A.2}$$

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

$$\arg\max_{\theta} \Pr(\mathcal{D} \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{I} \prod_{j=1}^{J_i} \Pr(a_{ij} \mid G_i, \theta).$$

[25] Equilibrium-based theories may have more of a role to play in the repeated setting, where agents have a chance to converge to equilibrium (although see Frey and Goldstone, 2013 for evidence against convergence in a repeated setting).

[26] This suggested value for τ may seem superficially similar to the value $\tau = 1.5$ suggested by Camerer et al. (2004) for Poisson-CH. However, they differ quite meaningfully, as $\tau = 1.12$ implies that 33% of the population are level-0, whereas $\tau = 1.5$ implies that only 22% are level-0.

[27] To those unfamiliar with Bayesian analysis, quantities such as $\Pr(\mathcal{D})$, $\Pr(G_i)$, and $\Pr(G_i \mid \theta)$ 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.

Table C.4
Datasets conditioned on various game features. The column headed "games" indicates how many games of the full dataset met the criterion, and the column headed "n" indicates how many observations each feature-based dataset contained. Observe that the game features are not all mutually exclusive, and so the "games" column does not sum to 142.

Name	Description	Games	n
D1	Weak dominance solvable in one round	2	748
D1s	Strict dominance solvable in one round	0	0
D2	Weak dominance solvable in two rounds	38	5058
D2s	Strict dominance solvable in two rounds	23	2000
DS	Weak dominance solvable	52	6470
DSs	Strict dominance solvable	35	3312
ND	Not dominance solvable	90	7393
PSNE1	Single Nash equilibrium, which is pure	51	4687
MSNE1	Single Nash equilibrium, which is mixed	21	1387
MULTI-EQM	Multiple Nash equilibria	70	7789

Appendix B. Posterior distribution derivation

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

$$\Pr(\theta \mid \mathcal{D}) = \frac{p_0(\theta)\Pr(\mathcal{D} \mid \theta)}{\Pr(\mathcal{D})}. \tag{B.1}$$

Substituting in Equation (A.2), which gave an expression for the likelihood of the dataset $\Pr(\mathcal{D} \mid \theta)$, we obtain

$$\Pr(\theta \mid \mathcal{D}) = \frac{p_0(\theta)\prod_{i=1}^{I}\prod_{j=1}^{J_i}\Pr(a_{ij} \mid G_i, \theta)\Pr(G_i)}{\Pr(\mathcal{D})}. \tag{B.2}$$

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

$$\Pr(\theta \mid \mathcal{D}) \propto p_0(\theta)\prod_{i=1}^{I}\prod_{j=1}^{J_i}\Pr(a_{ij} \mid G_i, \theta). \tag{B.3}$$

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.

Appendix C. Dataset composition

As we saw in the case of GH01, model performance was 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, and to evaluate model performance in each partition. In this appendix we describe such an analysis.

Overall, our datasets spanned 142 games. The vast majority of these games are matrix games, deliberately lacking inherent meaning in order to avoid framing effects.[28] 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.[29]

We thus constructed subsets of the full dataset based on their dominance solvability and the nature of their Nash equilibria, as described in Table C.4.[30] We computed cross-validated MLE fits for each model on each of the feature-based datasets of Table C.4. The results are summarized in Fig. C.8. 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 the majority of datasets; the exceptions were D1, D2, and D2s (but not DS); and MSNE1. Second, a majority of behavioral models significantly outperformed NEE in all but three datasets: D1, ND and MULTI-EQM. In these three datasets, the upper and lower

[28] Indeed, some studies (e.g., Rogers et al., 2009) even avoided focal payoffs like 0 and 100.

[29] There 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.

[30] As Table C.4 shows, there was some variance in the number of games and observations among the different partitions. The results presented in this appendix indicate that this variance was likely not a major determinant of our overall results.

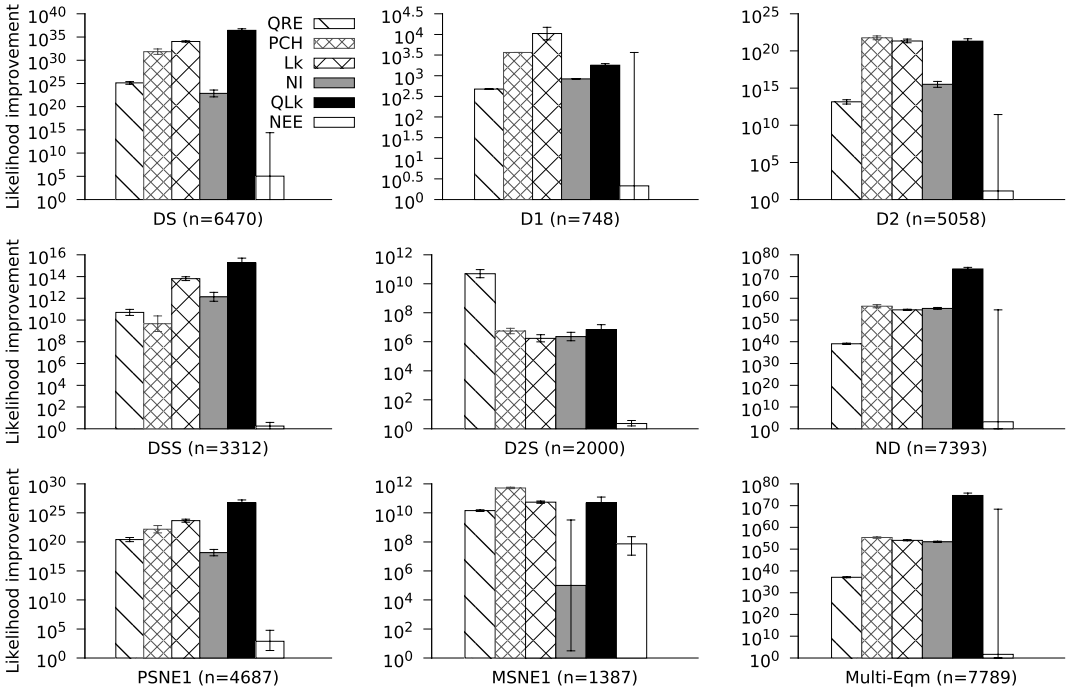

Fig. C.8. Average likelihood ratios of model predictions to random predictions, with 95% confidence intervals, on feature-based datasets. For NEE the main bar shows performance averaged over all equilibria and error bars show post-hoc upper and lower bounds on equilibrium performance.

bounds on NEE's performance contained the performance of either two or all three of the single-factor behavioral models (but not necessarily QLk). 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 90 games (and 4731 of the 7393 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 and DSS as the only exceptions; in D2S, QRE outperformed all other models, and in DSS QRE was significantly outperformed by Lk but not by Poisson-CH. 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 Lk and Poisson-CH that it outperforms on many datasets that mix game types. However, we observed no straightforward relationship between the different proportions of dominance-solvable and non-dominance-solvable games in a source dataset and the relative performance of Lk/Poisson-CH and QRE.

References

Altman, A., Bercovici-Boden, A., Tennenholtz, M., 2006. Learning in one-shot strategic form games. In: 17th European Conference on Machine Learning. ECML 2006, pp. 6–17.

Becker, T., Carter, M., Naeve, J., 2005. Experts Playing the Traveler's Dilemma. Working Paper. University of Hohenheim.

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer.

Breitmoser, Y., 2012. Strategic reasoning in p-beauty contests. Games Econ. Behav. 75 (2), 555–569.

Breitmoser, Y., Tan, J.H., Zizzo, D.J., 2014. On the beliefs off the path: equilibrium refinement due to quantal response and level-*k*. Games Econ. Behav. 86, 102–125.

Burchardi, K.B., Penczynski, S.P., 2014. Out of your mind: eliciting individual reasoning in one shot games. Games Econ. Behav. 84, 39–57.

Cabrera, S., Capra, C., Gómez, R., 2007. Behavior in one-shot traveler's dilemma games: model and experiments with advice. Spanish Econ. Rev. 9 (2), 129–152.

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

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

Camerer, C., Ho, T., Chong, J., 2004. A cognitive hierarchy model of games. Quart. J. Econ. 119 (3), 861–898.
Camerer, C., Hua Ho, T., 1999. Experience-weighted attraction learning in normal form games. Econometrica 67 (4), 827–874.
Camerer, C., Nunnari, S., Palfrey, T.R., 2016. Quantal response and nonequilibrium beliefs explain overbidding in maximum-value auctions. Games Econ. Behav. 98, 243–263.
Capraro, V., 2013. A model of human cooperation in social dilemmas. PloS One 8 (8), e72427.
Carvalho, D., 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. Rev. Econ. Design 16 (2–3), 215–250.
Chong, J., Camerer, C., Ho, T., 2005. Cognitive hierarchy: a limited thinking theory in games. In: Experimental Business Research, vol. III: Marketing, Accounting and Cognitive Perspectives, pp. 203–228.
Cooper, D., Van Huyck, J., 2003. Evidence on the equivalence of the strategic and extensive form representation of games. J. Econ. Theory 110 (2), 290–308.
Costa-Gomes, M., Crawford, V., 2006. Cognition and behavior in two-person guessing games: an experimental study. Amer. Econ. Rev. 96 (5), 1737–1768.
Costa-Gomes, M., Crawford, V., 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., Broseta, B., 2001. Cognition and behavior in normal-form games: an experimental study. Econometrica 69 (5), 1193–1235.
Costa-Gomes, M., Crawford, V., Iriberri, N., 2009. Comparing models of strategic thinking in Van Huyck, Battalio, and Beil's coordination games. J. Eur. Econ. Assoc. 7 (2–3), 365–376.
Costa-Gomes, M.A., Weizsäcker, G., 2008. Stated beliefs and play in normal-form games. Rev. Econ. Stud. 75 (3), 729–762.
Crawford, V., Iriberri, N., 2007a. Fatal attraction: salience, naivete, and sophistication in experimental "hide-and-seek" games. Amer. Econ. Rev. 97 (5), 1731–1750.
Crawford, V., 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.
Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: the 632+ bootstrap method. J. Amer. Statistical Assoc. 92 (438), 548–560.
Frey, S., Goldstone, R., 2011. Going with the group in a competitive game of iterated reasoning. In: 2011 Proceedings of the Cognitive Science Society, pp. 1912–1917.
Frey, S., Goldstone, R.L., 2013. Cyclic game dynamics driven by iterated reasoning. PloS One 8 (2), e56416.
Georganas, S., Healy, P.J., Weber, R.A., 2015. On the persistence of strategic sophistication. J. Econ. Theory 159, 369–400.
Gill, J., 2002. Bayesian Methods: A Social and Behavioral Sciences Approach. CRC Press.
Goeree, J.K., Holt, C.A., 2001. Ten little treasures of game theory and ten intuitive contradictions. Amer. Econ. Rev. 91 (5), 1402–1422.
Goeree, J.K., Holt, C.A., 2004. A model of noisy introspection. Games Econ. Behav. 46 (2), 365–382.
Goodie, A.S., Doshi, P., Young, D.L., 2012. Levels of theory-of-mind reasoning in competitive games. J. Behav. Decis. Mak. 25 (1), 95–108.
Hahn, P.R., Lum, K., Mela, C., 2010. A Semiparametric Model for Assessing Cognitive Hierarchy Theories of Beauty Contest Games. Working paper. Duke University.
Haile, P.A., Hortaçsu, A., Kosenok, G., 2008. On the empirical content of quantal response equilibrium. Amer. Econ. Rev. 98 (1), 180–200.
Hargreaves Heap, S., Rojo Arjona, D., Sugden, R., 2014. How portable is level-0 behavior? A test of level-k theory in games with non-neutral frames. Econometrica 82 (3), 1133–1151.
Haruvy, E., Stahl, D., 2007. Equilibrium selection and bounded rationality in symmetric normal-form games. J. Econ. Behav. Organ. 62 (1), 98–119.
Haruvy, E., Stahl, D., Wilson, P., 1999. Evidence for optimistic and pessimistic behavior in normal-form games. Econ. Letters 63 (3), 255–259.
Haruvy, E., Stahl, D., Wilson, P., 2001. Modeling and testing for heterogeneity in observed strategic behavior. Rev. Econ. Statist. 83 (1), 146–157.
Ho, T., Camerer, C., Weigelt, K., 1998. Iterated dominance and iterated best response in experimental "p-beauty contests". Amer. Econ. Rev. 88 (4), 947–969.
Jiang, A.X., Leyton-Brown, K., Bhat, N.A., 2011. Action-graph games. Games Econ. Behav. 71 (1), 141–173.
Kearns, M., Littman, M.L., Singh, S., 2001. Graphical models for game theory. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., pp. 253–260.
Koller, D., Milch, B., 2001. Multi-agent influence diagrams for representing and solving games. In: IJCAI, pp. 1027–1036.
McKelvey, R., Palfrey, T., 1995. Quantal response equilibria for normal form games. Games Econ. Behav. 10 (1), 6–38.
McKelvey, R., McLennan, A., Turocy, T., 2007. Gambit: software tools for game theory, version 0.2007.01.30.
Morgan, J., Sefton, M., 2002. An experimental investigation of unprofitable games. Games Econ. Behav. 40 (1), 123–146.
Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT Press.
Nagel, R., 1995. Unraveling in guessing games: an experimental study. Amer. Econ. Rev. 85 (5), 1313–1326.
Neal, R.M., 2001. Annealed importance sampling. Stat. Comput. 11 (2), 125–139.
Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Comput. J. 7 (4), 308–313.
Osborne, M.J., Rubinstein, A., 1998. Games with procedurally rational players. Amer. Econ. Rev. 88 (4), 834–847.
Patil, A., Huard, D., Fonnesbeck, C., 2010. PyMC: Bayesian stochastic modelling in python. J. Stat. Softw. 35 (1).
Polonio, L., Di Guida, S., Coricelli, G., 2015. Strategic sophistication and attention in games: an eye-tracking study. Games Econ. Behav. 94, 80–96.
Rey-Biel, P., 2009. Equilibrium play and best response to (stated) beliefs in normal form games. Games Econ. Behav. 65 (2), 572–585.
Robert, C.P., Casella, G., 2004. Monte Carlo Statistical Methods. Springer-Verlag.
Rogers, B.W., Palfrey, T.R., Camerer, C.F., 2009. Heterogeneous quantal response equilibrium and cognitive hierarchies. J. Econ. Theory 144 (4), 1440–1467.
Selten, R., Buchta, J., 1994. Experimental Sealed Bid First Price Auctions with Directly Observed Bid Functions. Discussion Paper B-270. University of Bonn.
Selten, R., Chmura, T., 2008. Stationary concepts for experimental 2×2-games. Amer. Econ. Rev. 98 (3), 938–966.
Shoham, Y., Leyton-Brown, K., 2008. Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations. Cambridge University Press.
Stahl, D., Haruvy, E., 2008. Level-n bounded rationality and dominated strategies in normal-form games. J. Econ. Behav. Organ. 66 (2), 226–232.
Stahl, D., Wilson, P., 1994. Experimental evidence on players' models of other players. J. Econ. Behav. Organ. 25 (3), 309–327.
Stahl, D., Wilson, P., 1995. On players' models of other players: theory and experimental evidence. Games Econ. Behav. 10 (1), 218–254.
Turocy, T., 2005. A dynamic homotopy interpretation of the logistic quantal response equilibrium correspondence. Games Econ. Behav. 51 (2), 243–263.
Von Neumann, J., 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 Econ. Behav. 44 (1), 145–171.
Witten, I.H., Frank, E., 2000. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann.
Wright, J.R., Leyton-Brown, K., 2010. Beyond equilibrium: predicting human behavior in normal-form games. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, pp. 901–907.
Wright, J.R., 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, vol. 2, pp. 921–928.
Wright, J.R., Leyton-Brown, K., 2014. Level-0 meta-models for predicting human behavior in games. In: Proceedings of the Fifteenth ACM Conference on Economics and Computation. EC'14, pp. 857–874.