MODEL SELECTION ACCURACY IN BEHAVIORAL GAME THEORY: A SIMULATION[†]

PAUL J. HEALY[*] AND HYOEUN PARK[**]

ABSTRACT. We simulate a horse race between several behavioral models of play in one-shot games. First, we find that many models can lead to identical predictions, making it impossible to select a unique winning model. This is largely avoided by comparing only two models. But even then we find that cross-validation sometimes fails to select the true model, often because models are be estimated to be noiseless but then fail to predict out-of-sample data. The Bayesian Information Criterion avoids this problem, though the inflexibility of its parameter penalty appears to cause poor performance in certain settings.

Keywords: Behavioral game theory; level-k; cognitive hierarchy; quantal responses

JEL Classification: C72; C92; D90; C52.

November 27, 2022

[†] The authors thank James Bland, Federico Echenique, Charlie Holt, Jaromir Kovářík Tom Palfrey, John Rehbeck, Ted Turocy, two anonymous referees and an associate editor, and the members of the OSU Theory/Experimental Reading Group for valuable comments.

[*] Dept. of Economics, Ohio State University, Columbus, Ohio 43210, U.S.A.; healy.52@osu.edu.

[**] Dept. of Economics, Ohio State University, Columbus, Ohio 43210, U.S.A.; park.2774@osu.edu.

I. Introduction

The behavioral game theory literature has produced several models of non-equilibrium play in one-shot games. Prominent examples include Quantal Response Equilibrium (McKelvey and Palfrey, 1995), the Level-k model (Stahl and Wilson, 1994; Nagel, 1995), and the Cognitive Hierarchy model (Camerer et al., 2004). The goal of each of these models is to explain and predict human behavior in strategic settings. Therefore, given a dataset of actual game play, it is natural to ask which model best explains the data we observe. To answer this, researchers often run a "horse race" of various models on their dataset, using a model selection criterion such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or one of many cross-validation methods to select a winning model. But how reliable is this exercise? Although these model selection criteria are known to have good large-sample properties, how well do they perform for a typical small-sample experiment dataset with only a handful of observations? Even in the starkest case where there is a true model that generated the data, would these various model selection methods correctly identify the data generating process (DGP) as the winning model?

To answer this question, we perform a model selection exercise on simulated data. We consider seven behavioral game theory models and, for each model, generate a dataset of 3,000 simulated subjects who play twelve 3×3 one-shot games according to that model. Thus, each model serves as the DGP for one dataset. For each subject in each dataset we perform a model-selection exercise to see which of the seven models is selected as the winning model for that subject. If the model selection exercise is successful then the DGP should be selected as the winning model for the vast majority of subjects. If not, then this raises serious concerns about the validity of the model selection exercise.

Our first result is that the models based on iterated best-reply make very similar predictions, causing the model selection exercise to fail or to be inconclusive because several models perform equally well.[1]

We then explore possible solutions to this problem. First, we note that these models are better separated when game payoffs are smaller. This is because most models predict that noise increases as payoff differences shrink, and it is this noise in behavior that often distinguishes similar models. Unfortunately, we find that model selection is not substantially improved by this change. And in some cases the performance is strictly worse.

Our next solution is to restrict ourselves to comparing only two competing models, and ensuring that those two models are well separated. Of our seven models, six are based on levels of iterated best response, while Quantal Response Equilibrium (QRE) is

[1] García-Pola et al. (2020) make a similar observation in the context of centipede games.

the only model that is not. Thus, we compare each levels-based model against QRE. For those datasets generated by a levels-based model, we compare that model against QRE. Then, for the QRE dataset we perform six separate model comparisons, each featuring QRE against one of the six levels-based models.

Model selection performance improves substantially when comparing only two models. However, when using cross validation methods for selecting the winning model we still find a handful of unexpected anomalies in which the wrong model wins for a substantial fraction of subjects. We find that one of the biggest problems with cross validation is that it is prone to "infinite penalties." This happens when a model is flexible enough to fit the training data perfectly, causing it to be estimated as a noiseless, deterministic model. If this deterministic model ever makes an incorrect prediction then it is penalized with a zero likelihood (or, a negative-infinity log likelihood) and will not be selected. If this happens to the model that generated the data, then a "wrong" model will necessarily be selected. The AIC and BIC avoid the infinite penalty problem, and thus perform more reliably than cross validation in our simulated exercise. We also identify two other kinds of failures that can occur with cross validation, both of which are again avoided by switching to the AIC or BIC; see Section IV for details.

The twelve 3 × 3 games we study were adapted from past work and not necessarily optimized to discriminate between these seven models. In Section V we ask if model selection performance would be improved by using optimized games. We find that this is the case, though the failures of cross validation identified in our original simulation still appear with these optimized games. As an additional result, we show that having subjects play twelve copies of a single game does not significantly affect the frequency of these anomalies.

Increasing the number of games that each subject plays should also improve the model selection performance. But here the infinite penalty problems that arise with cross validation can actually cause it to perform better for smaller numbers of games. This occurs because the non-DGP model is more likely to have a negative-infinity failure with fewer games, giving the DGP an extra advantage when the number of games is small. Again, the Bayesian and Akaike Information Criteria avoid this problem and behave more predictably as the number of games is increased. We confirm these patterns by varying the number of games used in our simulation.

Finally, a larger strategy space should allow for better model discernment, simply because there are more available actions on which the models can differ. We test this by switching to a simulation of two-person guessing games (Costa-Gomes and Crawford, 2006), which have continuum strategy spaces. We then restrict play to either a coarse grid or a fine grid over those strategy spaces, where the fine grid has ten times as many available strategies as the coarse grid. With cross validation we find that moving to the fine strategy space offers relatively small improvements, though they are significant for some of the data generating processes.

With the BIC and AIC, however, results are more mixed and there are even two data generating processes for which performance is significantly worse in the fine strategy space. We conjecture that the BIC and AIC perform worse than the cross-validation methods here because their overfitting penalties are too inflexible. Ideally the penalty should vary as the strategy space changes, since this affects the propensity to overfit the data. For example, it is much harder for a model to overfit data with a fine strategy space, and so a smaller penalty may be appropriate. But the BIC and AIC penalties do not adjust in this way, so their performance may vary across settings.

At first blush our results suggest that researchers should (1) avoid similar models, and (2) use either the AIC or BIC to avoid problems with cross validation. But the AIC and BIC don't always behave as expected, either, as the penalties they impose may be too rigid to apply uniformly across settings. Thus, our recommendation is that researchers interested in performing model selection on their data should first run a simulation similar to ours to verify that the criteria they plan to use will work well and feature the expected comparative statics on the set of games they are studying.

Our simulations assume there always exists a true data generating process for any dataset. This is necessary for us to declare unequivocally whether the "right" model was selected for a given subject. Presumably, this provides an upper bound on the performance of any model selection methodology: If a given criterion cannot identify the right model when one exists, then it seems unlikely to identify the "better" model (suitably defined) when none are correct.[2] We leave the many complications of this question for future research, and focus here on only the simplest case in which a correct model exists.

We compare seven behavioral game theory models: six levels-based models in the spirit of Level-*k* (Nagel, 1995; Stahl and Wilson, 1994; Camerer et al., 2004; Costa-Gomes and Crawford, 2006), and Quantal Response Equilibrium (McKelvey and Palfrey, 1995). But the point of our exercise is not specific to these models. Our main finding is that similar models can cause problems, and that even when we avoid similar models there can be peculiar situations where the wrong model wins for various reasons. Thus, simulating the model selection exercise before applying it to real data can help identify such failures.

[2] For example, one could imagine simulating a population of subjects in which 80% conform to one model and 20% to another, and then asking whether the former model is selected when forcing only one model to fit the entire population. This is different from our setting because we identify a winning model for each individual subject.

Similarly, the particular failures and comparative statics we observe may be specific to the games we chose—which were adapted from Stahl and Wilson (1995) and Costa-Gomes and Crawford (2006)—but again the general lesson that model selection should be treated with caution is likely not. In all cases it is worthwhile to simulate the exercise first to validate its use in any given setting.

To our knowledge, ours is the first simulation of model selection with one-shot plays of games. Salmon (2001) performs a similar exercise by comparing various learning models on simulated datasets in which subjects play a single game repeatedly. His criterion for success, however, differs in that he only asks whether estimates of model parameters correctly or incorrectly rule out reinforcement learning versus belief learning. He finds that models that should not fit a given data generating process often pass the test, indicating serious type II (false positive) failures. In a similar vein, Feltovich (2000) finds that, on actual experimental data, whether a reinforcement model or a belief learning model fits better depends not only on the game, but also on the success criterion used. Thus, model selection appears very sensitive in the learning domain, and prone to overfitting problems.

Carbone and Hey (1994a) test the discrimination between models of individual choice—such as expected utility (EU) and several non-EU models—using simulated data. They apply both the Akaike Information Criterion and the index developed in Carbone and Hey (1994b), take EU as the null hypothesis, and ask how often it is rejected in favor of another model. They find that when the data are truly generated by an EU maximizer then the non-EU theories are rarely selected. But when the data are from a non-EU subject then EU is rejected in favor of multiple non-EU theories. Thus, the non-EU theories appear hard to discriminate. Carbone (1997) extends this analysis to both binary choice data and rank-order list data, and also finds that the "wrong" model can be selected quite often, particularly when the data are generated from a non-EU model.[3]

In the domain of one-shot games, three examples of model selection exercises are given by Breitmoser (2012), Wright and Leyton-Brown (2017), and García-Pola et al. (2020). All three perform model selection exercises on actual experimental data. Breitmoser (2012) compares level-k, logistic level-K, QRE, and noisy introspection (Goeree and Holt, 2004) in p-beauty contests using a Vuong test, which is essentially identical to the BIC in terms of penalizing parameters. He finds that quantal response and noisy introspection explain most play better than the level-k model. Wright and Leyton-Brown (2017) compare level-k, quantal level-k, cognitive hierarchy, noisy introspection, and QRE to each other and to a model based on Nash equilibrium by using 10 rounds of 10-fold cross-validation. They find support for the quantal level-k model above all

[3] Jakusch (2013) applies this method to financial data and explores how variations in the assumed error structure can affect the results.

others. In neither paper is the reliability of the model selection criterion verified, so the propensity for false positives is unclear.

Fudenberg et al. (2020) and Fudenberg et al. (2022) attempt to quantify how prone a given model is to overfitting. One characteristic of an overfitting model is that it will have low prediction error regardless of the true data generating process (DGP). Fudenberg et al. (2020) imagine randomly generating many different DGPs, and then measuring how well the model fits each of these randomly-generated DGPs. If the model is prone to overfitting then its average error will be small; this represents a lack of *restrictiveness* for that model. In contrast, the *completeness* of a model (Fudenberg et al., 2022) is a measure of the model's error compared to the *true* DGP. An ideal model is one that has a large value of completeness (it fits the actual DGP well) and also a large value of restrictiveness (it wouldn't fit other DGPs well). One can think of this approach as defining (or informing) an intrinsic preference over models, which can be evaluated without any data. In contrast, cross validation, AIC, and BIC all provide methods for model selection when a dataset is given, and vary in how they penalize models that are prone to overfitting.

García-Pola et al. (2020) compare the predictions of eleven behavioral models in centipede games played as normal-form games. They note that much of the past literature on centipede games used payoffs that failed to properly distinguish these models. To correct for this, they design 16 novel centipede games that vary only in their payoffs and verify theoretically that the models are sufficiently discriminated.[4] Instead of picking a single winning model for each subject, García-Pola et al. (2020) estimate a mixture model that allows each subject to play according to multiple models, each with some estimated probability; see McLachlan and Peel (2000) for details. Their results show that a mixture of the Level-k and QRE models fits the data well.

One promising direction for model selection is to use non-choice data to further distinguish between models. Johnson et al. (2002), Costa-Gomes et al. (2001), Costa-Gomes and Crawford (2006) and others gather data on what information subjects view while making decisions. For example, if a subject does not collect enough information to calculate their opponent's best responses then that subject presumably cannot be a level-2 subject who is responding to their opponent's best response. Chen et al. (2018) similarly use eye-tracking to see how subjects mentally "calculate their strategy" in a novel spatial beauty contest and use this data to infer how many levels of reasoning a given subject must be using. To our knowledge, however, these methods have only been used to help fit parameters within a given model, but have not been used to help select a winning model from several contenders.

[4] They employ both a spike-logit and a spike-uniform error structure; see Section II for a description of spike-logit errors.

II. The Models

A two-player game is given by $(\mathscr{S}_1, \mathscr{S}_2, u_1, u_2)$, where $\mathscr{S}_i$ is the (finite) strategy space of player $i \in \{1,2\}$ and $u_i : \mathscr{S}_1 \times \mathscr{S}_2 \to \mathbb{R}$ specifies the payoff to i for each strategy profile $(s_1, s_2) \in \mathscr{S}_1 \times \mathscr{S}_2$.[5]

A mixed strategy for player i is given by $\sigma_i \in \Delta(\mathscr{S}_i)$, where $\Delta(\mathscr{S}_i)$ is the space of all distributions over $\mathscr{S}_i$. For example, in a 3×3 game $\sigma_i = (1/3, 1/3, 1/3)$ represents a uniform mixture over player i's three strategies. Following a common abuse of notation, let

$$u_i(\sigma_1, \sigma_2) = \sum_{(s_1, s_2)} \sigma_1(s_1)\sigma_2(s_2)u_i(s_1, s_2);$$

$u_i(s_1, \sigma_2)$ and $u_i(\sigma_1, s_2)$ are defined similarly.

For any σ_j let

$$BRS_i(\sigma_j) = \arg\max_{s_i} u_i(s_i, \sigma_j)$$

be the set of best responses for i against σ_j. Then define $BR_i(\sigma_j)$ to be the mixed strategy that uniformly randomizes over $BRS_i(\sigma_j)$. Formally,

$$BR_i(\sigma_j)(s_i) = \begin{cases} \frac{1}{\#BRS_i(\sigma_j)} & \text{if } s_i \in BRS_i(\sigma_j) \\ 0 & \text{if } s_i \notin BRS_i(\sigma_j). \end{cases}$$

Each player plays a set of games $\mathscr{G} = \{1, \ldots, G\}$; when needed we index a player's strategies by $g \in \mathscr{G}$.[6]

Many models use logistic response rather than best response. Formally, for any subset of strategies $\mathscr{S}_i' \subseteq \mathscr{S}_i$, belief σ_j, and precision parameter $\lambda \geq 0$, the logistic response (restricted to $\mathscr{S}_i'$) of player i is given by the distribution

$$LR_i(\sigma_j | \lambda, \mathscr{S}_i')(s_i) = \begin{cases} \frac{\exp(\lambda u_i(s_i, \sigma_j))}{\sum_{s_i' \in \mathscr{S}_i'} \exp(\lambda u_i(s_i', \sigma_j))} & \text{if } s_i \in \mathscr{S}_i' \\ 0 & \text{if } s_i \notin \mathscr{S}_i'. \end{cases}$$

With this notation we can now describe the seven behavioral game theory models that we include in our model selection exercise.

Level-k

Following the previous literature (Nagel, 1995; Stahl and Wilson, 1994, 1995; Costa-Gomes and Crawford, 2006, etc.) we anchor the Level-*k* model on a "level 0" type that uniformly randomizes over all possible actions. Level 1 then best responds to level 0,

[5] All of the models we consider do not differentiate between (objective) pecuniary payoffs and (subjective) utility indices, so we take u_i to represent pecuniary payoffs.

[6] Both u_i and $\mathscr{S}_i$ depend on g as well, but we ignore this in our notation. We also use i and j to index both individuals in the experiment and player roles within a game.

level 2 best responds to level 1, and so on. But this model makes deterministic predictions (generically), which often leads to datasets with a zero likelihood. Thus, following Costa-Gomes and Crawford (2006), we add a "spike-logit" error structure in which the player plays according to their best response with probability $1-\epsilon$, and plays a logistic response with probability ϵ. The logistic distribution has precision parameter λ, so the model has three total parameters: k, ϵ, and λ.

We study two variants of the model that differ slightly in how they model these ϵ-probability "trembles" from the best resopnse strategy. In the "double counting" version of the model (denoted by "LK Double" or simply LKD) the logistic distribution puts weight on all strategies, meaning that trembles can include best responses. In the "single counting" version (LKS) the logistic distribution excludes best-response strategies, meaning all trembles are strictly suboptimal.

In either version of the model the player does not believe their opponent will tremble.

To formalize the model we first define the best response strategy for each level inductively, with

$$\sigma_i^{\mathrm{LK}}(s_i|0) = \frac{1}{\#\mathscr{S}_i} \ \forall s_i \in \mathscr{S}_i, \text{ and}$$
$$\sigma_i^{\mathrm{LK}}(\cdot|k) = BR_i\left(\sigma_j^{\mathrm{LK}}(\cdot|k-1)\right) \ \ \forall k \in \{1,2,\ldots\}.$$

For brevity, let

$$BRS_i^k = BRS_i(\sigma_j^{\mathrm{LK}}(\cdot|k-1)),$$

and for any $\mathscr{S}_i' \subseteq \mathscr{S}_i$ let $\mathbb{1}_{\mathscr{S}_i'}(s_i)$ be the indicator function for $s_i \in \mathscr{S}_i'$. The LKD and LKS strategies are then defined (for each $k>0$) by

$$\begin{aligned}\sigma_i^{\mathrm{LKD}}(s_i|k,\lambda,\epsilon) = {} & \mathbb{1}_{BRS_i^k}(s_i)\cdot(1-\epsilon)\cdot BR_i\left(\sigma_j^{\mathrm{LK}}(\cdot|k-1)\right)(s_i) \\ & + \epsilon\cdot LR_i\left(\sigma_j^{\mathrm{LK}}(\cdot|k-1)|\lambda,\mathscr{S}_i\right)(s_i)\end{aligned}$$

and

$$\begin{aligned}\sigma_i^{\mathrm{LKS}}(s_i|k,\lambda,\epsilon) = {} & \mathbb{1}_{BRS_i^k}(s_i)\cdot(1-\epsilon)\cdot BR_i\left(\sigma_j^{\mathrm{LK}}(\cdot|k-1)\right)(s_i) \\ & + \left(1-\mathbb{1}_{BRS_i^k}(s_i)\right)\cdot\epsilon\cdot LR_i\left(\sigma_j^{\mathrm{LK}}(\cdot|k-1)|\lambda,\mathscr{S}_i\setminus BRS_i^k\right)(s_i).\end{aligned}$$

Again, the difference between these is whether the logistic-response tremble probabilities are assigned to every strategy (LKD), or only those that are not best responses (LKS).

Finally, we add a Nash type, denoted by $k=\infty$. All of our games have a unique (though possibly mixed) Nash equilibrium, so we can define the best response of the Nash type directly from the fixed point given by

$$\sigma_i^{\mathrm{LK}}(\cdot|\infty) \in \arg\max_{\sigma_i} u_i(\sigma_i, \sigma_j^{\mathrm{LK}}(\cdot|\infty)).$$

The LKD and LKS models then add logistic noise to this Nash strategy. Letting

$$BRS_i^{\infty} = BRS_i(\sigma_j^{\mathrm{LK}}(\cdot|\infty))$$

be the support of $\sigma_i^{\mathrm{LK}}(\cdot|\infty)$, we have

$$\begin{aligned}\sigma_i^{\mathrm{LKS}}(s_i|\infty,\lambda,\epsilon) &= \mathbb{1}_{BRS_i^{\infty}}(s_i)\cdot(1-\epsilon)\cdot\sigma_i^{\mathrm{LK}}(s_i|\infty)\\ &\quad+\left(1-\mathbb{1}_{BRS_i^{\infty}}(s_i)\right)\cdot\epsilon\cdot LR_i\left(\sigma_j^{\mathrm{LK}}(\cdot|\infty)|\lambda,\mathscr{S}_i\setminus BRS_i^{\infty}\right)(s_i)\end{aligned}$$

and

$$\begin{aligned}\sigma_i^{\mathrm{LKD}}(s_i|\infty,\lambda,\epsilon) &= \mathbb{1}_{BRS_i^{\infty}}(s_i)\cdot(1-\epsilon)\cdot\sigma_i^{\mathrm{LK}}(s_i|\infty)\\ &\quad+\epsilon\cdot LR_i\left(\sigma_j^{\mathrm{LK}}(\cdot|\infty)|\lambda,\mathscr{S}_i\right)(s_i).\end{aligned}$$

For our simulations, the grid of parameter values is $k \in \{1,2,3\}$, $\lambda \in \{0.01, 0.05, 0.16, 0.56, 1, 1.4, 1.8, 2.2, 2.3, 2.4, 2.45, 2.55, 2.6, 2.7, 3.2, 4.2, 4.7, 5.2, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 60\}$, and $\epsilon \in \{0, 0.05, 0.1, 0.15, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$.[7]

Poisson Cognitive Hierarchy

The Poisson Cognitive Hierarchy (PCH) model is similar to the Level-k model, but features different subjective beliefs about the types of their opponents. Following Camerer et al. (2004), the model assumes that type k's belief over her opponents follows a truncated Poisson distribution over lower types.

Specifically, if $f(k|\tau) = \exp(-\tau)\frac{\tau^k}{k!}$ is the Poisson distribution with parameter τ, then a player of type k believes that each type $k' \in \{0,1,\ldots,k-1\}$ occurs with frequency $g(k'|k,\tau) = f(k'|\tau)/\sum_{k''=0}^{k'-1} f(k''|\tau)$, and types $k' \geq k$ never occur ($g(k'|k,\tau) = 0$).[8]

[7] The parameters $\lambda = 20$ and $\lambda = 60$ give almost same strategies, so there is rarely a meaningful difference between them. And for our main games the strategies calculated by MATLAB are identical between $\lambda = 60$ and $\lambda = \infty$. For the small payoff games described later, we use $\lambda \in \{0.01, 0.05, 0.16, 0.56, 1, 1.4, 1.8, 2.2, 2.3, 2.4, 2.45, 2.55, 2.6, 2.7, 3.2, 4.2, 4.7, 5.2, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 6000\}$. This is because we need larger λ to have the same prediction as $\lambda = \infty$, and to have consistency in estimation with QRE.

[8] It is possible that players believe others perform more steps of reasoning than themselves, but since they cannot conceive of what strategy such a person would play they lump these players together with the level-0 behavior.

Define the noiseless strategies inductively by

$$
\begin{aligned}
\sigma_i^{\mathrm{PCH}}(s_i|0,\tau) &= \frac{1}{\#\mathscr{S}_i}\ \forall s_i \in \mathscr{S}_i,\\
\gamma_i^{k,\tau} &= \sum_{k'=0}^{k-1} g(k'|k,\tau)\,\sigma_j^{\mathrm{PCH}}(\cdot|k',\tau)\ \ \forall k \in \{1,2,\ldots\}, \text{and}\\
\sigma_i^{\mathrm{PCH}}(\cdot|k,\tau) &= BR_i\left(\gamma_i^{k,\tau}\right)\ \ \forall k \in \{1,2,\ldots\}.
\end{aligned}
$$

Here, $\gamma_i^{k,\tau}$ is type k's overall belief about the actions of their opponent, to which they best respond.

As in the LK model we add spike-logit noise to these strategies, which can either be of the single-counting (PCHS) or double-counting (PCHD) form. Define the best response sets by

$$BRS_i^{k,\tau} = BRS_i(\gamma_i^{k,\tau}).$$

The model's strategies are then defined by

$$
\begin{aligned}
\sigma_i^{\mathrm{PCHS}}(s_i|k,\tau,\lambda,\epsilon) = {}& \mathbb{1}_{BRS_i^{k,\tau}}(s_i)\cdot(1-\epsilon)\cdot BR_i\left(\gamma_i^{k,\tau}\right)(s_i)\\
&+\left(1-\mathbb{1}_{BRS_i^{k,\tau}}(s_i)\right)\cdot\epsilon\cdot LR_i\left(\gamma_i^{k,\tau}|\lambda,\mathscr{S}_i\setminus BRS_i^{k,\tau}\right)(s_i)
\end{aligned}
$$

and

$$
\begin{aligned}
\sigma_i^{\mathrm{PCHD}}(s_i|k,\tau,\lambda,\epsilon) = {}& \mathbb{1}_{BRS_i^{k,\tau}}(s_i)\cdot(1-\epsilon)\cdot BR_i\left(\gamma_i^{k,\tau}\right)(s_i)\\
&+\epsilon\cdot LR_i\left(\gamma_i^{k,\tau}|\lambda,\mathscr{S}_i\right)(s_i).
\end{aligned}
$$

The grid of parameter values used in our simulation is the same as in the Level-k model, with the grid for the added parameter τ being $\tau \in \{0.4, 0.6, 0.8, 1.2, 1.6, 2, 2.4, 2.8, 3.2, 3.6, 4\}$.

Hierarchical Quantal Response

The hierarchical quantal response model (HQR) is similar to the level-k models (LKS and LKD), but with two modifications: First, players use logistic response with no "spike" on the pure best response. Second, players correctly believe that their opponent plays a logistic response, rather than a pure best response.

Formally, the model is defined inductively by

$$
\begin{aligned}
\sigma_i^{\mathrm{QR}}(s_i|0,\lambda) &= \frac{1}{\#\mathscr{S}_i}\ \forall s_i \in \mathscr{S}_i, \text{and}\\
\sigma_i^{\mathrm{QR}}(\cdot|k,\lambda) &= LR_i\left(\sigma_j^{\mathrm{QR}}(\cdot|k-1,\lambda)|\lambda,\mathscr{S}_i\right)\ \forall k \in \{1,2,\ldots\}.
\end{aligned}
$$

This model has only two parameters (k and λ) and the grids used for them are the same as in the Level-k and PCH models.

HQR is a special case of Noisy Introspection (Goeree and Holt, 2004). Noisy introspection allows for different values of λ and an infinite hierarchy of levels, while HQR requires a common λ and only considers $k \in \{1,2,3\}$.

Quantal Level-k

Stahl and Wilson (1994) suggest the quantal level-k model (QLK), which takes only the first two levels of the HQR model but allows level 2 to have a different precision parameter than level 1, and for level 2 to have incorrect beliefs about the precision parameter of level 1.

Let $\vec{\lambda} = (\lambda^1, \lambda^2, \lambda^{1(2)})$ be the vector of precision parameters, where λ^1 is level 1's actual precision, λ^2 is level 2's actual precision, and $\lambda^{1(2)}$ is level 2's (degenerate) belief about level 1's precision. For each $s_i \in \mathscr{S}_i$ define

$$\begin{aligned}
\sigma_i^{QLK}(s_i|0,\vec{\lambda}) &= \frac{1}{\#\mathscr{S}_i},\\
\sigma_i^{QLK}(\cdot|1,\vec{\lambda}) &= LR_i\left(\sigma_j^{\mathrm{QLK}}(\cdot|0,\vec{\lambda})|\lambda^1,\mathscr{S}_i\right),\\
\gamma_i^{1(2)}(\cdot|\vec{\lambda}) &= LR_i\left(\sigma_j^{\mathrm{QLK}}(\cdot|0,\vec{\lambda})|\lambda^{1(2)},\mathscr{S}_i\right), \text{and}\\
\sigma_i^{QLK}(\cdot|2,\vec{\lambda}) &= LR_i\left(\gamma_i^{1(2)}(\cdot|\vec{\lambda})|\lambda^2,\mathscr{S}_i\right),
\end{aligned}$$

where $\gamma_i^{1(2)}$ is level 2's (possibly incorrect) belief about level 1's strategies. Like HQR, QLK is also a special case of the Noisy Introspection model of Goeree and Holt (2004), but limited to only two levels.

This model has four total parameters. Following Stahl and Wilson (1994), we only allow $k \in \{1,2\}$. The grid used for all three lambda parameters is the same as the grid used for λ in the models described above.

Quantal Response Equilibrium

Finally, we consider the logit quantal response equilibrium (QRE) of McKelvey and Palfrey (1995). In this equilibrium model all players have correct beliefs but apply logistic response (over the entire strategy space) rather than perfect best response. Thus, it is a fixed point of the logistic response function, rather than the best response correspondence.

Formally, a (logistic) QRE in a two-player game is a mixed strategy profile such that, for each $i \in \{1,2\}$,

$$\sigma_i^{\mathrm{QRE}}(\cdot|\lambda) = LR_i\left(\sigma_j^{\mathrm{QRE}}(\cdot|\lambda)|\lambda,\mathscr{S}_i\right)$$

A game can have multiple quantal response equilibria; in that case we limit attention only to the equilibrium on the principal branch, which exists for every λ (McKelvey and Palfrey, 1995, Theorem 3). This is found using simple homotopy methods, starting at the centroid ($\lambda = 0$) and numerically tracing the branch to the desired value of λ (see Turocy, 2005, e.g.).

The QRE model has only one parameter, λ. We apply the same grid as in the other models.

III. The Model Selection Exercise

We perform a simulated model-selection exercise using the seven behavioral models described in Section II. We have no human subjects; instead, we simulate players playing 12 two-player, symmetric 3×3 games adapted from Stahl and Wilson (1995). Six games have pure strategy Nash equilibria and the other six have totally mixed Nash equilibria. The payoff matrices are presented in Table I.

For each model we generate a dataset of 3,000 simulated subjects who each play the 12 games according to that model. We refer to this model as the DGP for that dataset. Each simulated subject is assigned randomly-drawn parameters for that model. Parameter values are independently drawn from each grid with a distribution that's roughly uniform. The grids are not equally spaced, however; parameter values closer to those estimated in prior research are sampled more frequently. Each simulated subject then plays all 12 games in accordance with the model and their randomly-drawn parameter values. If the model prescribes a mixed strategy for a subject then the computer randomly draws one pure strategy from the specified mixed strategy distribution, and this becomes the subject's chosen pure strategy. A dataset therefore consists of a 3,000-by-12 matrix of actions. Each model has one such dataset for which it is the DGP.

Next, for each dataset we perform a model selection exercise. This simulates a researcher who observes the dataset but does not know its true DGP. We perform model selection at the individual level on all 3,000 simulated subjects. Specifically, we estimate different parameters for each subject individually and use the four model selection criteria to select the winning model for each subject. We use the same parameter grid for estimation as was used to draw the DGP parameters, so that the "true" parameters are available when estimating the model. If the model selection procedure is accurate then the DGP will be selected as the winning model in the vast majority of the 3,000 subjects. Our primary metric for success will therefore be the fraction of simulated subjects for whom the DGP was selected.

There are several approaches to model selection. We compare four popular methods: Leave-One-Out Cross Validation (LOOCV), 2-Fold Cross Validation (2FCV), Bayesian

Table I. Game Payoffs: Baseline Simulation

Game 1	T	M	B
T	25	30	100
M	40	45	65
B	31	0	40

Game 2	T	M	B
T	30	50	100
M	40	45	10
B	35	60	0

Game 3	T	M	B
T	10	100	40
M	0	70	50
B	20	50	60

Game 4	T	M	B
T	30	100	50
M	40	0	90
B	50	75	29

Game 5	T	M	B
T	30	100	22
M	35	0	45
B	51	50	20

Game 6	T	M	B
T	40	15	70
M	22	80	0
B	30	100	55

Game 7	T	M	B
T	25	30	100
M	40	0	65
B	31	45	40

Game 8	T	M	B
T	10	100	40
M	0	70	60
B	20	50	50

Game 9	T	M	B
T	39	15	70
M	40	80	0
B	30	100	55

Game 10	T	M	B
T	30	50	100
M	40	60	10
B	35	45	0

Game 11	T	M	B
T	30	100	22
M	35	0	20
B	51	50	45

Game 12	T	M	B
T	40	80	60
M	23	25	0
B	30	100	55

Information Criterion (BIC), and Akaike Information Criterion (AIC). Each of these methods selects a winning model by comparing the likelihood of the subject's actions under each model, but with a correction to avoid overfitting. In the BIC and AIC the likelihood values are reduced by an explicit penalty that is increasing in the number of parameters. The cross-validation methods instead estimate each model's parameter values using one subset of the 12 games (the training set) and then evaluate each model's likelihood on the complementary set of games (the testing set). These methods therefore correct for overfitting by using out-of-sample likelihoods as the criterion.

Formally, let $s_i = (s_i^1, \dots, s_i^{12})$ be subject i's observed strategies in the 12 games. Suppose model M has r parameters, denoted by the vector θ. Fix a set of games $A \subseteq \mathcal{G}$ and a simulated subject i. For any value of θ the likelihood of observing a vector of strategies $s_i^A = (s_i^g)_{g \in A}$ under model M is given by

$$L^{\mathrm{M}}(s_i^A|\theta) = \prod_{g \in A} \sigma_i^{\mathrm{M}}(s_i^g|\theta).$$

The MLE estimate of θ for subject i is $\hat{\theta}_i^A = \arg\max_\theta L^{\mathrm{M}}(s_i^A|\theta)$. When all games are used ($A = \mathscr{G}$) we drop A from the notation.

The BIC and AIC are based on MLE estimates over all games in $\mathscr{G}$, with a penalty for the number of parameters r. Formally, the BIC for subject i (with 12 observations) is given by

$$BIC^{\mathrm{M}}(s_i) = \ln(L^{\mathrm{M}}(s_i|\hat{\theta}_i)) - \frac{\ln 12}{2} r.$$

The AIC is

$$AIC^{\mathrm{M}}(s_i) = \ln(L^{\mathrm{M}}(s_i|\hat{\theta}_i)) - r.$$

The model with the highest $BIC^M(s_i)$ or $AIC^M(s_i)$ is then declared the winning model for subject i.

The AIC provides an estimate of the expected information loss when using an incorrect model, as measured by the Kullback-Leibler divergence between the incorrect model and the true model (Akaike, 1974). The BIC (Schwarz, 1978) was instead designed to be an estimate that is proportional to a Bayesian researcher's posterior belief about the model being the true model, which becomes prior-independent when the sample size grows large.[9] For regression models, the probability of selecting the true model (when it exists) goes to one for the BIC, but not for the AIC. Small-sample properties of the AIC and BIC are known for linear models, where it is generally accepted that the AIC over-rewards models with many parameters (Hurvich and Tsai, 1989, e.g.). Our models are neither regression models nor linear models, so it is less clear which criterion is more desirable. Yang (2005) writes that, for finite-dimensional models, the consensus view is that BIC is preferred over AIC for model selection. For our simulation with 12 games the BIC does provides a stronger penalty on the number of parameters, since $(\ln 12)/2 \approx 1.242$ is greater than one.

Two-Fold Cross Validation (2FCV) instead takes $s_i = (s_i^1, \ldots, s_i^{12})$ and randomly splits it into two vectors of six observations each. Each subject's split may be different. Denote the split for subject i by A_i and B_i, where $A_i \cup B_i = \mathscr{G}$. Here, $s_i^{A_i}$ is the training data and $s_i^{B_i}$ is the testing data. Parameters are estimated as $\hat{\theta}_i^{A_i} = \arg\max_\theta L^M(s_i^{A_i}|\theta)$ using the training set A_i, and then the log-likelihood value used for model selection is

$$2FCV^M(s_i) = \ln\left(L^M(s_i^{B_i}|\hat{\theta}_i^{A_i})\right),$$

which uses the testing set B_i. The winning model for subject i is that with the highest value of $2FCV^M(s_i)$.

[9] The result for the AIC requires model errors to be Gaussian. The result for the BIC assumes the models come from an exponential family. Neither assumption is satisfied by all of the models we test, so we cannot interpret them as estimators. Regardless, it is common to use these criteria even when the assumptions are violated.

LOOCV performs 12 cross-validation estimations, one for each observation. Let $\iota \in \{1,\dots,12\}$ index the twelve estimations. In each step the ιth game alone is used as the testing set, so let $s_i^{B_\iota} = (s_i^\iota)$. The other eleven observations form the training set $s_i^{A_\iota}$. Estimated parameters for model M in step ι are given by $\hat{\theta}_i^{A_\iota} = \arg\max_\theta L^M(s_i^{A_\iota}|\theta)$, and the likelihood value on the ιth testing set is $L^M(s_i^{B_\iota}|\hat{\theta}_i^{A_\iota})$. The log-likelihood values of the 12 estimations are then averaged. The model with the highest value of

$$LOOCV^M(s_i) = \frac{1}{12}\sum_{\iota=1}^{12} \ln\left(L^M(s_i^{B_\iota}|\hat{\theta}_i^{A_\iota})\right)$$

is declared the winning model for subject i.

Assuming only some differentiability conditions, Stone (1977) shows that the LOOCV criterion converges to the AIC as the sample size grows large. Shao (1997) provides a similar asymptotic equivalence between the BIC and 2FCV, but only for linear models.[10]

For each model we count the number of subjects for which that model wins. In some cases multiple models may be selected as the winner. For example, LK Double and LK Single become identical when $\hat{\epsilon} = 0$; if that is the case then these two models necessarily must give the same likelihood values. We report separately how often each model wins uniquely and how often each model wins in a tie with other models.

When simulating the 3,000 subjects for each model we draw parameters from each grid using a method that generates a roughly uniform distribution over the parameter grid. Recall, however, that the grids are finer in regions where parameter values are more plausible, meaning we oversample parameter values that are closer to those estimated in past work. Also, our procedure for randomly generating parameter vectors unintentionally led to an additional oversampling of lower values of ϵ and λ on the grids, and a significant (but small) negative correlation between λ and ϵ. This is true for all models except QRE. In Appendix C we describe these deviations from uniformity. We then re-run our simulation using parameters that are truly uniform on the grids and without correlation. Qualitatively all results are the same, indicating that our results are robust to small changes in the distributions of parameters.

We will occasionally be interested in what parameter estimates we observe for each subject. With cross-validation the parameters are estimated using subsamples of the data, which is not an efficient estimate. Thus, after a winning model is selected we re-estimate its parameter values using the entire sample, giving an overall parameter estimate of $\hat{\theta}_i = \arg\max_\theta L^M(s_i|\theta)$.

[10] Specifically, asymptotic equivalence is achieved for linear models when $|A_i|$ (the size of the training set) is $n/(\ln(n)-1)$, which is roughly eight games for $n = 12$. Since our focus is neither on asymptotics nor on linear models, we only examine 2FCV with equal splits of the data since this is by far the most common practice.

DGP	# Parameters	LOOCV	2FCV	BIC	AIC
LK Double	3	0.83	0.83	0.7	0.93
LK Single	3	44.93	35.2	57.6	59.1
PCH Double	4	3.7	8.23	12.13	14.07
PCH Single	4	15.13	10.43	31.33	35.97
QLK	4	1.73	2.27	4.8	5.1
HQR	2	2.43	1.5	92.5	92.87
QRE	1	19.13	44.5	94.33	92.13

TABLE II. Frequency with which each data generating process wins uniquely.

Recall that our primary metric of success for each method is how frequently that method chooses the true DGP model as the winning model. Hurvich and Tsai (1989) and Rao et al. (2008) use similar metrics when analyzing their simulated data.

IV. MAIN RESULTS

Table II shows for each DGP the percentage of observations in which that DGP is selected as the unique winning model. Our first result is that model selection exercise largely fails, regardless of the criterion used.[11] The only cases where the DGP wins more than 75% of the time occur when the DGP has only one or two parameters and the AIC or BIC are used. For the cross-validation methods the average accuracy across all models is 24.7% for LOOCV and 34.7% for 2FCV. The BIC and AIC have slightly higher average accuracy (44.4% and 45.3%, respectively), but with much higher variance across models.

If we view each model as a hypothesis, then the DGP is a null hypothesis being tested against six alternative hypotheses. The typical significance threshold of 5% for hypothesis tests would mean that the entries in Table II should be greater than 95%. In fact none are, so one interpretation is that none of these model selection exercises provides a test that has an adequate significance threshold.

Indeed, some perform very poorly. LK Double is correctly selected in less than 1% of the subjects, while HQR and QRE (which have the fewest parameters) are correctly selected in over 90% of subjects. It is apparent that the parameter punishments in the AIC and BIC are excessive here, favoring only those models with the fewest parameters.

Lesson 1. No model selection criterion guarantees high accuracy across all models.

A major reason for the model selection exercise to fail is because several of the models are similar, and in fact can become identical under certain parameter values. This leads

[11] On the other hand, we can reject the null of completely random model selection. Under that hypothesis 95% of the winning percentages would be in the interval [13.0%, 15.5%].

DGP	# Parameters	LOOCV	2FCV	BIC	AIC
LK Double	3	92.97	93.1	0.03	0
LK Single	3	23.13	20.33	0	0
PCH Double	4	71.23	71.07	25.77	25.77
PCH Single	4	20.7	17.5	8.47	8.47
QLK	4	84.93	63.83	0.1	0.07
HQR	2	84.93	89.6	0	0
QRE	1	0	0.07	0	0

TABLE III. Frequency with which each DGP wins in a tie with another model.

to a high frequency of ties between models. Table III reports how frequently models win in a tie.[12] From this table we see that with cross validation ties can happen very frequently, often for a large majority of subjects.

To illustrate how ties happen, suppose the DGP for a given subject is LK Double with $k = 2$ and ϵ small. Then it is reasonably likely that this subject will play a perfect level-2 best response (without noise) in all 12 games. In that case both LK Double and LK Single will have parameter estimates $\hat{k} = 2$ and $\hat{\epsilon} = 0$, leading to 100% likelihood for both models. PCH Double and PCH Single will also achieve 100% likelihood on this data by having $\hat{\tau}$ large enough, while QLK and HQR will achieve 100% likelihood by having $\hat{\lambda}$, $\hat{\lambda}^{1(2)}$, and $\hat{\lambda}^2$ all large enough. In other words, these models are not identifiable for this subject's data.

The BIC and AIC largely avoid these ties because the parameter penalty serves as a tie-breaking rule when two models generate the same likelihood but have different numbers of parameters. Whether this tie-breaking rule correctly selects the DGP depends entirely on whether the DGP happens to have the fewest parameters among the models that tie.

Table IV provides finer detail for this result, showing the winning frequency of every model for every DGP. It separates cases where each model wins uniquely ("solo") from those cases where it ties either with the DGP (and possibly other models; "tie-DGP"), or where it ties with other non-DGP models ("tie-others"). Here we focus only on LOOCV; tables for the other methods appear in the appendix. We can see from the main diagonal of the table that almost all of the six levels-based models win in a tie far more often than they win solo. In fact, four of them almost never win solo. If instead we ask whether the DGP wins more frequently than any other model (regardless of its absolute winning frequency), then only one model—LK Single—succeeds according to this metric. The high incidence of ties indicates that these games do not provide adequate discrimination between these levels-based models.

[12] The DGP's total win frequency is the sum of the values from Tables II and III.

DGP	win type	LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
LK Double	solo	0.83	0.73	0.4	0.37	0.63	0.47	2.6
	tie-DGP	N/A	92.83	62.13	62.1	62.1	92.93	0.5
	tie-others	92.97	0.07	0.77	0.83	0.17	0.17	0
LK Single	solo	3.43	44.93	2.63	4.23	2.2	3.97	10.23
	tie-DGP	21.87	N/A	16.63	17.9	16.63	21.83	1.57
	tie-others	2.87	23.13	1.83	0.2	2.73	2.93	0
PCH Double	solo	3.1	1.57	3.7	1.7	7.4	4.47	4.6
	tie-DGP	47.07	47.03	N/A	71.13	48.13	47.1	1.33
	tie-others	1.53	0.07	71.23	0.07	0.7	2.1	0
PCH Single	solo	5.4	19.87	4.6	15.13	6.3	8	14.17
	tie-DGP	11.93	13.17	19.47	N/A	12.53	11.93	0.6
	tie-others	3.73	0.1	1.97	20.7	2.67	3.5	0
QLK	solo	1.23	2.03	0.63	0.57	1.73	1.07	6.2
	tie-DGP	82.6	82.47	84.03	84.03	N/A	83.23	0.4
	tie-others	1.13	0.43	0.2	0.47	84.93	0.97	0
HQR	solo	1.13	0.63	0.87	0.63	1.07	2.43	5.9
	tie-DGP	84	83.1	54.9	54.83	55.9	N/A	0.17
	tie-others	0.17	0.43	1.97	2.23	0.03	84.93	0
QRE	solo	42.53	0.9	0.53	0.67	1.6	1.73	19.13
	tie-DGP	0	0	0	0	0	0	N/A
	tie-others	31.43	31.47	1.13	1.23	0.83	0.83	0

TABLE IV. For each DGP (row) and model (column), the percentage of subjects for which the model wins uniquely ("solo"), wins in a tie with the DGP (and possibly other models; "tie-DGP"), and wins in a tie with other non-DGP models ("tie-others").

We conclude that a researcher interested in comparing various models should be wary of the identification problem that arises when comparing similar models. And one simple way to identify whether two models are similar is to perform a simulation exercise similar to ours to check how frequently the models cannot be discriminated.

Lesson 2. When selecting among similar models it is important to verify how frequently they can be discriminated.

One solution to the non-identification problem is to alter the structure of the games in a way that increases model identification. The main difference between the levels-based models is how noise is modeled, and games in which strategies are closer to indifferent (according to the beliefs given by the model) should lead to more noise and therefore greater discriminatory power. Since the level of noise depends on the size of the payoffs, we should observe noisier play when payoffs are lower. Therefore we scale our 3×3 game payoffs by 1/100 to increase the noise in play. Noise also enters into beliefs in the QLK, QR, and QRE models, so a reduction in payoffs can actually change the ordinal ranking of strategies for these models, but not for the Level-k or PCH models. For both reasons we expect that lower payoffs will improve model identification.

DGP	win type	LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
LK Double	solo	8.07	5.7	4.73	2.7	3.5	5.17	10
	tie-DGP	N/A	53.97	38.2	36.6	36.93	54.4	1.83
	tie-others	56.3	0.83	1.23	2.03	1.8	1.8	0
LK Single	solo	9.27	27.37	5.57	5.43	4.73	6.9	14.03
	tie-DGP	15.47	N/A	10.97	15.93	10.97	15.47	1.43
	tie-others	3.67	20.43	3.57	0.37	2.53	2.53	0
PCH Double	solo	7.77	4.77	4.87	5.23	5.63	6.67	14.57
	tie-DGP	27.77	25.67	N/A	45.23	25.67	25.73	1.2
	tie-others	1.33	0.73	47.33	0.53	1.5	2.3	0
PCH Single	solo	9.3	15.27	4.83	11.9	6.63	8.47	15.47
	tie-DGP	7.77	11.97	15.53	N/A	7.77	7.77	0.8
	tie-others	6.1	0.07	5.1	19.73	2.9	3.5	0
QLK	solo	8.33	7.57	3.2	5.77	5.87	14.17	17.23
	tie-DGP	25.57	24.93	25.2	24.93	N/A	28.1	0.33
	tie-others	3.9	0.53	6.77	5.63	28.33	2.43	0
HQR	solo	7.3	7	3.43	4.5	3.7	14.93	15.17
	tie-DGP	38.4	36.7	24.77	24.4	27.33	N/A	0
	tie-others	1.6	0.97	1.7	1.27	0.2	41.1	0
QRE	solo	24.07	6.57	4.1	6.1	6.3	11.4	25.63
	tie-DGP	0	0	0	0	0	0	N/A
	tie-others	13.93	12.7	1.23	0.77	1.67	1.87	0

TABLE V. A replication of Table IV when game payoffs are scaled by 1/100.

Table V provides the win and tie frequencies of each DGP and each model when the payoffs in the games are divided by 100. Although identification is somewhat improved we still see from the main diagonal of the table that most DGPs win much more often in ties than solo. Other patterns—such as single-counting error out-performing double-counting error and LK Double beating QRE—remain similar to those found with the larger payoff scale.

Overall we conclude that a structural change in the games is not enough to achieve satisfactory identification with our set of models.

Lesson 3. Structural changes to the games may not be enough to overcome the identification problems that arise when comparing similar models.

Our second solution to the identification problem is simply to omit similar models. In our original exercise (Table IV) we see that the six levels-based models frequently tie with other levels-based models, but when QRE is the DGP it never ties with any other model. Interestingly, however, it still only wins for 19.13% of subjects. The LK Double model wins 74% of the time, with 42.5% of wins being solo and another 31.4% being ties with LK Single. So it is not immediately clear that omitting similar models will fix the model selection problem.

DGP\EST	LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
LK Double	95.13						4.87
LK Single		76.35					23.65
PCH Double			77.52				22.48
PCH Single				*45.73*			54.27
QLK					89.27		10.73
HQR						89.68	10.32
QRE QRE wins:	*24.32*	*48.92*	96.23	96.8	96.03	93.43	
other wins:	75.68	51.08	3.77	3.2	3.97	6.57	

TABLE VI. LOOCV winning frequency of each model versus only QRE in the 3 ×3 games with the original payoffs.

To test this, we perform six pairwise model-selection exercises on each levels-based model versus QRE. Specifically, for each of the six levels-based models (LKD, LKS, PCHD, PCHS, QLK, and HQR) we set that model as the DGP and compare it head-to-head against QRE using LOOCV. Then we set QRE as the DGP and run it head-to-head against each of the other six models.

The results are shown in Table VI. For the first six rows the main diagonal gives the success rate of the DGP while the last column gives the failure rate, which is the frequency with which QRE wins. In the bottom row QRE is the DGP. The top number in each cell is the success rate while the bottom is the failure rate. Now we see much stronger identification, with the DGP winning in well over 75% of instances for most models. Though the more stringent requirement of a 95% success rate (based on a 5% hypothesis testing significance level) is only satisfied in four of the 21 comparisons. Tests of QRE as the DGP perform the best: it wins more than 93% of the time against all models except the LK models.

There are, however, three anomalous cases in which the DGP does not overwhelmingly win the two-horse horse race. These are italicized in Table VI and we explain each below.

Anomaly 1: PCH Single vs. QRE

Consider the case where PCH Single is the DGP. Note that its extra parameter (τ) allows it to have a wider variety of best response profiles, compared to the LK models. For example, if we look at the vector of best responses across the 12 games, LK level 2 has a unique vector of best responses, while PCH level 2 has six possible vectors that can be best responses, depending on τ. This flexibility results in over-fitting problems that can be quite severe. In particular, it makes it more likely that the model is estimated to be a deterministic model (for example, with $\hat{\epsilon} = 0$ or $\hat{\lambda} = \infty$), which gets an infinite penalty

DGP:	LK Double	LK Single	PCH Double	PCH Single	QLK	HQR
# Losses:	138	682	636	1605	316	307
% with $-\infty$:	71.0%	91.5%	81.0%	84.5%	62.7%	52.8%

TABLE VII. Of those subjects for whom the DGP loses to QRE, the percentage that have (approximately) $-\infty$ likelihood values.

if that model's predictions turn out to be wrong in the testing data. We now detail three scenarios that can generate this problem.

First, suppose that a subject happens to play the best response profile of some PCH parameter combination (k,τ) for all 11 training games. In this case, $\hat{\epsilon} = 0$ for both the double and single counting models. If the subject does not play the best response in the testing game then the log likelihood value will be $-\infty$. If this happens even once out of the 12 cross-validation folds then the PCH model cannot win.

Second, suppose that the subject always plays a non-best response for the 11 training games for some (k,τ). In this case $\hat{\epsilon} = 1$. If the subject then chooses the best response in the testing game then the log likelihood value will also be $-\infty$.

Finally, suppose that a subject in the training games plays the best responses for some (k,τ) in n_1 $(0 \le n_1 < 11)$ games and the "second best" response for (k,τ) in the other $11 - n_1$ games (meaning, the strategy with the second-highest expected payoff). In this case, $\hat{\lambda} = \infty$ in the single counting model. This is because the $11 - n_1$ plays must be considered trembles, but the error structure precludes playing the true best response when trembling, so an infinite λ correctly predicts that the subject will always play the "second best" response. If the subject actually plays the "third best" response for (k,τ) in the testing game then the PCH-Single likelihood is $-\infty$. For the double counting model this cannot happen; $\hat{\lambda}$ would remain finite, preventing the model from making deterministic predictions.

In our simulation PCH Single loses to QRE in 1,605 out of 3,000 subjects (excluding ties). We find that 84.49% of these failures are explained by one of the above three cases.

In fact, $-\infty$ likelihood problems are the leading cause of losses for all six levels-based DGPs. Table VII shows that, while PCH Single losses are far more common, the *fraction* of losses caused by a $-\infty$ likelihood is high for the other models as well.

At which parameter values does this anomaly occur most frequently? Although it is observed to some degree at every parameter vector, it is most frequent when the subject's true ϵ is close to one. Given the single-counting error structure, this means the subject very often trembles away from the best response action. This makes it very likely that $\hat{\epsilon} = 1$, as in the second case described above.

The second-most frequent occurrence of this anomaly is when ϵ is close to zero. Here the estimated parameter is often $\hat{\epsilon} = 0$, giving rise to the first case described above. The third case arises roughly equally across parameter values. We don't find that the true values of λ, τ, or k have a significant impact on the frequency of this anomaly. See the online appendix for details of these results.

Lesson 4. When a DGP can be estimated to be deterministic, cross validation methods can suffer from "infinite penalty" failures.

Anomaly 2: QRE vs. Both LK Models

Consider the case where QRE is the DGP and is competing against either LK model (the first two columns of the bottom row in Table VI). Since our baseline game payoffs are relatively large the QRE DGP often generates strategies consistent with Nash equilibrium, even for modest levels of λ. In that case the LK models can exactly predict this behavior since they include a Nash type. For example, in Game 6 if we exclude ties then LK Double beats QRE in 87.64% of cases. Of those, the LK Double model estimates the player to be the Nash type 94.79% of the time. And 80.11% of these are estimated to be a "noiseless" Nash type. By this we mean either that $\hat{\epsilon} = 0$ so that trembles never happen or, for the case of LK Double, $\hat{\lambda}$ is large enough so that "trembles" become noiseless best replies.

Now, one might expect that QRE should also be able to perfectly imitate noiseless Nash play by having a large estimated $\hat{\lambda}$. But recall that six of the games have a totally mixed Nash equilibrium, and suppose the realized strategy of a subject in one game has a relatively low (but positive) Nash equilibrium probability. In that case a lower $\hat{\lambda}$ actually gives a higher likelihood for QRE since the realized strategy is more likely to come from trembles than from Nash play. This is what we see: despite perfect Nash play being very frequent in the dataset, the QRE noise parameter $\hat{\lambda}$ is estimated to be small (weakly less than one) for 85.8% of subjects. And this causes it to predict worse than the LK models on the testing data, because the testing data is most likely to be a pure-strategy or high-probability Nash equilibrium action.

Anomaly 3: QRE vs. LK Double

In addition to the previous anomaly, LK Double gains a second advantage over QRE that is unique to its double-counting error structure.

First, we establish that in the LK Double model with a Nash type, even if all realized strategies are consistent with Nash equilibrium it still might be that the estimated weight on trembles is positive ($\hat{\epsilon} > 0$). To see why, suppose games 1–11 are used as the testing set and each of player i's realized strategies are all in the support of the Nash equilibrium. Let s_g be the realized strategy in each game g, suppose games 1–6 have a pure strategy Nash equilibrium, and games 7–12 have a totally mixed equilibrium. Then the log-likelihood for the LK Double's Nash type in game g can be written simply as $\log[(1-\epsilon)\cdot NE(s_g)+\epsilon LR(s_g)]$, where $NE(s_g)$ is the Nash probability of the realized strategy s_g and $LR(s_g)$ is the logistic response probability assuming the opponent plays Nash noiselessly. If game g has a pure strategy equilibrium then $NE(s_g)=1$ and $LR(s_g)<1$, and if g has a totally mixed equilibrium then $NE(s_g)<1$ and $LR(s_g)=1/3$ because the opponent's mixed Nash strategy makes the player indifferent between all three strategies. Thus, the likelihood function becomes

$$\sum_{g=1}^{6}\log[(1-\epsilon)+\epsilon LR(s_g)]+\sum_{g=7}^{12}\log[(1-\epsilon)NE(s_g)+\epsilon/3].$$

The first sum is always decreasing in ϵ since $LR(s_g)<1$ for all s_g. If $NE(s_g)>1/3$ for all games 7 through 11 then the second sum would also be decreasing in ϵ, giving $\hat{\epsilon}=0$. But if $NE(s_g)<1/3$ for some games then the second sum is increasing, so an interior solution for $\hat{\epsilon}$ may obtain. In the LK Single model, however, the LR and 1/3 terms are replaced with zeros, giving $\hat{\epsilon}=0$ as the only possible solution.

Now, suppose the testing game ($g=12$) has a totally mixed equilibrium and $NE(s_{12})<1/3$. For the QRE model the likelihood will be close to $NE(s_{12})$, especially for larger $\hat{\lambda}$. But for LK Double the likelihood will be

$$(1-\varepsilon)\,NE(s_{12})+\varepsilon\,\frac{1}{3}\geq NE(s_{12}),$$

with strict inequality of $\varepsilon>0$. It is in these situations that LK Double gains an extra advantage over QRE due to its double-counting error structure, and this explains why it beats QRE even more frequently than LK Single.

How often does this happen? It requires that (1) the subject plays the pure-strategy Nash equilibrium in the five or six games in the training set that have a pure-strategy equilibrium, and (2) $NE(s_g)<1/3$. Since QRE converges quickly to Nash equilibrium in games with a pure strategy equilibrium but not for games with mixed equilibria, we find that both conditions are reasonably likely for moderate ranges of λ. For example, for $\lambda=0.56$ the subject has a roughly 99% chance of playing a pure-strategy Nash equilibrium in each game that has one, but in games with mixed-strategy equilibria they can play a strategy such that $NE(s_g)<1/3$ with as much as 50% probability.[13]

In our simulation 74% of our subjects satisfy these two conditions in at least one of the twelve iterations of the LOOCV procedure. And in the other 11 iterations the two models either tie or else LK Double wins for the reason described in Anomaly 2. Thus,

[13] For example, in game $g=4$ we have $NE=(0.53,0.17,0.30)$, so two strategies have $NE(s_g)<1/3$. But the QRE probabilities when $\lambda=0.56$ are $(0.51,0.19,0.30)$, which gives a 49% chance of playing an s_g with $NE(s_g)<1/3$.

DGP\EST	LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
LK Double	97.53						2.47
LK Single		93.07					6.93
PCH Double			92.17				7.83
PCH Single				76.7			23.3
QLK					93.5		6.5
HQR						93.97	6.03
QRE QRE wins:	98.57	97.7	98.37	97.6	98.87	97.1	
Model wins:	1.43	2.3	1.63	2.4	1.13	2.9	

TABLE VIII. BIC winning frequency of each model versus only QRE in the 3 ×3 games with the original payoffs

LK Double ultimately wins over QRE in these cases. In the online appendix we show that this is true across almost all values of λ except the smallest, which is exactly where Nash play becomes infrequent.

Lesson 5. When a model (such as Nash) generates a low-probability action, many models will estimate higher levels of noise, thus reducing fit on the higher-probability actions.

BIC vs. LOOCV

While cross-validation is often seen as the gold standard for model selection, we've seen that there can be cases where it fails. In particular, if a very flexible model's parameters are estimated such that its predictions become deterministic, and if those deterministic predictions prove to be wrong, then the model is infinitely penalized. This infinite penalty is impossible under the BIC and AIC since the models are never asked to make out-of-sample predictions. Here we show that this in fact makes the BIC superior to LOOCV in our domain.

Table VIII shows the result of the binary model selection exercise using BIC instead of LOOCV.[14] Indeed, it performs better in every case, compared to LOOCV (Table VI). All winning frequencies are now above 75%, and seven of the 12 are above 95%. Simply put, the BIC provides fewer false negatives because its overfitting penalty is generally less extreme.

For example, if QRE is the DGP and we simply compare unadjusted likelihood values then LK Double beats QRE for 60.1% of subjects. This reflects the LK Double model overfitting the data. After the BIC adjustment, however, LK Double wins in less than 2% of cases. Similarly, LK Single beats QRE in 48.1% of cases using unadjusted likelihoods, but in only 2.3% of cases after the BIC adjustment.

[14] Due to numerical limitations we regard log-likelihood values less than -7 (averaged across all 12 cross-validation folds) as being $-\infty$.

In Section V, however, we will show that the BIC penalty may be less optimal as we move to different types of games or different numbers of games. So although the BIC outperforms LOOCV for our base games, it should still be used with caution because its penalty only depends on the number of parameters and not on other factors that might also affect the degree of overfitting.

V. Robustness and Sensitivity

In this section we explore whether the failures of LOOCV that we have identified will persist as the number of games changes, when the size of the strategy space changes, and when the games' payoffs are chosen to maximize the distance between the various models' predictions.

Choosing Payoffs to Maximize Model Differentiation

The 12 games we chose were adapted from Stahl and Wilson (1995). Some games are taken directly from the original paper, while others are modified to be as "close" to the original games as possible while ensuring six have a unique pure strategy equilibrium and six have a unique mixed strategy equilibrium. But these twelve games were not specifically designed to maximize the discrimination between the seven models we consider.[15] Here we ask whether model selection performance would improve if the games' payoffs were chosen to separate maximally the various models' predictions. Doing so should reduce the number of ties between models. But what's less clear is whether the anomalies we identify with LOOCV would be mitigated with these optimized games.

To that end, we first need a metric of how well a given game separates two models. Specifically, if model M_a generates a predicted mixed strategy distribution $\sigma_i^{M_a}(\cdot|\theta_a)$ for game g at parameter vector θ_a and model M_b generates $\sigma_i^{M_b}(\cdot|\theta_b)$ for game g at parameter vector θ_b, then we need a measure of distance between $\sigma_i^{M_a}(\cdot|\theta_a)$ and $\sigma_i^{M_b}(\cdot|\theta_b)$. And we need to account for the fact that this distance will depend on which parameter values are used.

For our distance metric we opt for the simple Euclidean (or, L^2) distance given by

$$D_g^2(M_a,M_b;\theta_a,\theta_b) = \frac{1}{\sqrt{2}}\sqrt{\sum_{s_i\in S_i}\left(\sigma_i^{M_a}(s_i|\theta_a) - \sigma_i^{M_b}(s_i|\theta_b)\right)^2},$$

which ranges from zero to one. Then, to average this measure across possible parameter values, we generate 1,000 randomly-drawn parameter vectors of the form $\theta =$

[15] Indeed, most of these models didn't exist in 1995 when the original experiment was run.

$(\theta_{LKd},\ldots,\theta_{QRE})$ and calculate the average value of $D^2_g(M_a,M_b;\theta_a,\theta_b)$ over these 1,000 θ values.[16] We denote this "expected discrimination" for game g by $E_\theta D^2_g(M_a,M_b)$.

Many papers use the Kullback-Leibler (KL) divergence to measure the distance between two models' predictions; see for example El-Gamal and Palfrey (1996) or Balietti et al. (2021).[17] One difficulty, however, is that if one of the models puts zero (or very low) probability on a certain action while the other does not then the KL divergence becomes infinite. If this happens for even one parameter vector then the expected discrimination cannot be calculated.[18] We opted for the simpler L^2 distance to avoid this complication.

Next we need to assess whether $E_\theta D^2_g(M_a,M_b)$ is relatively large or small. To do that, we generate 1,000 random 3×3 games, calculate the expected discrimination for each, and see how the expected discrimination for our 12 games compare to the distribution of 1,000 random games. If our games are near the top of this distribution then arguably they are nearly optimal for discriminating between these models.

But we need to make sure the randomly-drawn games are roughly comparable to the original 12 in terms of the magnitude of payoffs. We therefore draw the payoffs for the 1,000 random games with replacement from the set of payoffs used in the original twelve games.[19] The rationale for this is that the average payoffs in the random games should be, on average, the same as the original 12 games. This mimics a researcher with a fixed budget and a uniform prior about the play of subjects. And it also makes sure that the random games don't perform better or worse than the original games simply because of differences in payoff magnitudes.

Fix two models M_a and M_b. To compare the expected discrimination of our original 12 games to that of the 1,000 randomly generated games, we calculate for each original game $g\in\{1,\ldots,12\}$ the percentage of randomly-drawn games that have a weakly lower expected discrimination than game g. Formally, if $F_{ab}(\cdot)$ is the empirical cdf of expected discrimination for the 1,000 random games, we calculate $F_{ab}(E_\theta D^2_g(M_a,M_b))$. If game g is unusually good at differentiating M_a and M_b then this value will be close to one.

[16] The 1,000 parameter vectors are drawn (with replacement) from the same support as in the main simulation process and with the same distribution.

[17] Balietti et al. (2021) actually measure the distance between multiple models simultaneously across multiple games. We focus here on discriminating only pairs of models, given our conclusions above, and do so for each game separately.

[18] Taking the median KL distance among the 1,000 parameter vectors, rather than the mean, doesn't entirely solve the problem since infinite distances are quite common and can affect the median. Conclusions then become very sensitive to how often these infinite distances occur.

[19] Specifically, for every payoff entry in a random game, we randomly and uniformly select one cell from one of the original 12 games and use that payoff entry. This process is done with replacement, and the payoff entries in the random game need not all come from the same original game. Recall that the games are symmetric, so each cell has only one payoff entry.

Model 1\Model 2	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
LK Double	0.985	0.996	0.992	0.997	0.998	0.915
LK Single		0.994	0.988	0.995	0.97	0.92
PCH Double			0.998	0.98	0.975	0.94
PCH Single				0.991	0.973	0.929
QLK					0.975	0.924
HQR						0.914

TABLE IX. $\max_{g\in\{1,\ldots,12\}} F_{ab}(E_\theta D_g^2(M_a,M_b))$ for each pair of models M_a and M_b.

DGP\EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
PCH Single					74.55			25.45
QRE	QRE wins:	51.83	52.37					
	Model wins:	48.17	47.63					

TABLE X. LOOCV winning frequency of the DGP for 12 games with the highest $D_2(m_1,m_2;\theta_1,\theta_2)$. Only models with anomalies are shown.

Figures III and IV in Appendix B show the entire distribution F_{ab} and the values of $F_{ab}(E_\theta D_g^2(M_a,M_b))$ for each of our 12 original games. For the 12-game experiment to be successful at discriminating M_a from M_b we may not need all 12 games to score highly. Even if only one or two games are successful at discriminating then the model selection exercise can still succeed. Thus, in Table IX we report the maximum of these 12 values for each pair of models. If the games we chose were randomly selected then we would expect this maximum to be 12/13 ≈ 0.923. For most pairs of models our best game ranks far higher than this, though for the comparisons against QRE our best game appears no better on average than the best of any random 12 games.[20] In other words, our games achieve unusually high discrimination among the levels-based models, but have only average discrimination when comparing those models to QRE.[21]

Again, achieving good theoretical separation between models doesn't necessarily mean that the anomalies we identified with LOOCV will disappear. To test that, we re-run our simulation, replacing our original 12 games with the 12 random games that achieved the best discrimination between a given pair of models. Table X shows that the winning frequency of the DGP does improve, but the anomalies do not disappear. For the PCH Single anomaly the winning frequency increases from 44.07% to 74.55%, while for the

[20] Figures III and IV reveal that our games are clearly different from a random set of 12 games—for example the minimum is almost never as low as 1/13 ≈ 0.077—, so this result is only a statement about the best game, not all 12 games.

[21] Figures III and IV also show that which game is best varies from one pair of models to the next. And that for each of our games there is at least one pair of models where that game is important in achieving separation. Thus, each game plays a role in achieving discrimination between the seven models.

DGP\EST	LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
PCH Single				77.5			22.5
QRE QRE wins:	51.23	51.18					
Model wins:	48.77	48.82					

TABLE XI. LOOCV winning frequency of the DGP for 12 repetitions of the one game with the highest $D_2(m_1, m_2; \theta_1, \theta_2)$. Only models with anomalies are shown.

two QRE anomalies it increases from 24.32% to 51.83% and from 48.92% to 52.37%, respectively. Thus, optimizing the games based on model discrimination appears to help, but does not completely eliminate the problems associated with LOOCV.

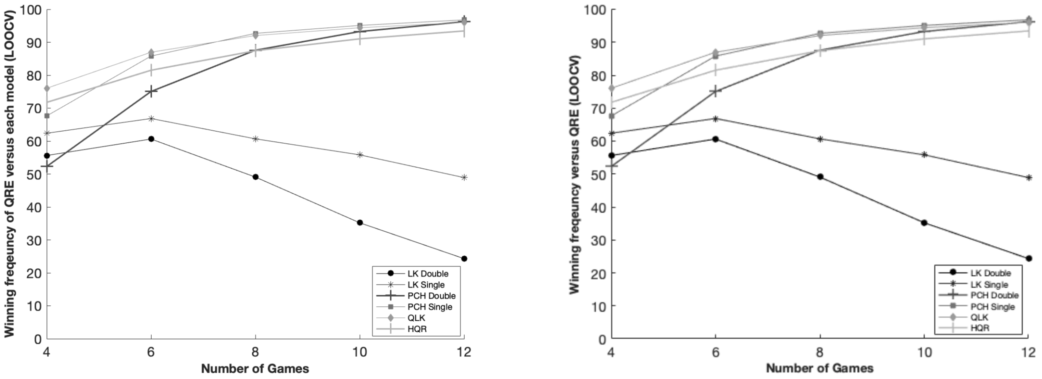
Finally, we ask whether model selection would be improved and anomalies reduced if subjects played the same game 12 times, rather than playing 12 different games. To answer this, we re-run our simulation, this time replacing our original 12 games with 12 copies of the one random game that achieved the best separation. We did this for each pair of models. The results appear in Table XI. The winning frequencies are very similar to those in Table X, suggesting that very little is gained by having subjects play the same game repeatedly.

Number of Games

In the limit, model selection should become more accurate as the number of games is increased. For a researcher facing time and budget constraints, a natural question is how much improvement is achieved by adding an additional game. To address this, we re-run our simulation on subsets of the original 12 games. Specifically, for each DGP we have all 3,000 subjects play a random subset of four games, compare that DGP against QRE (or QRE against each of the other models), and look at the fraction for whom the DGP is correctly identified. We then repeat this exercise for random subsets of six, eight, and ten games.

The results for LOOCV are shown in Figure I. The graphs show that the success rates of the models (panel a) are surprisingly insensitive to the number of games. There are two interesting phenomena to notice here: First, the winning frequency of PCH Double actually decreases in the number of games. Second, when the DGP is QRE (panel b) its winning frequencies against LK Double and LK Single both decrease in the number of games.

For the first case of PCH Double, there are two major forces. The first comes from deviations from the best response strategy of a given level. When the number of games is small it is reasonably likely that a PCH Double player will perfectly best respond in every game. This is particularly likely with the double-counting error structure, since



(a) (b)

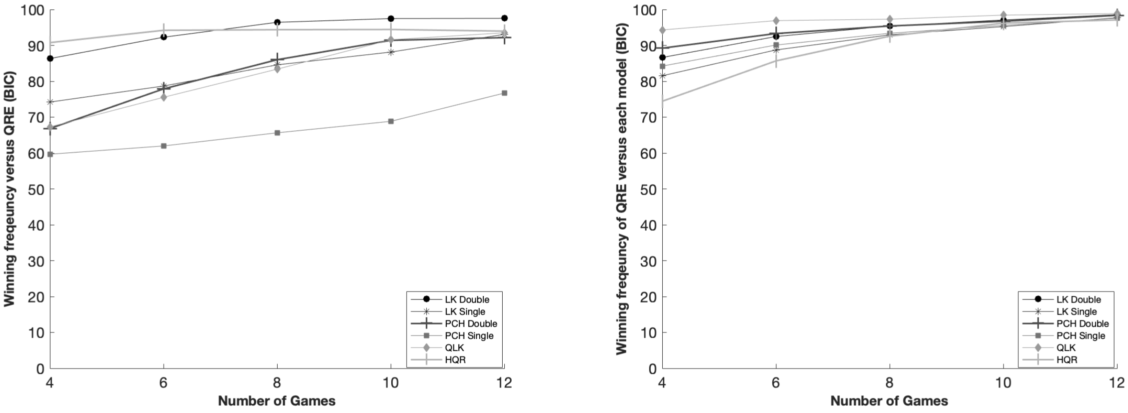
FIGURE I. Winning frequencies when using LOOCV. Panel (a): Winning frequencies of each model as the DGP. Panel (b): Winning frequency of QRE as the DGP against each model.

even when a player trembles they may still play the best response strategy. When all realized strategies are consistent with the best responses then the PCH Double model gets a likelihood value of one (with $\hat{\epsilon} = 0$) and so it beats QRE. But as the number of games grows this becomes less likely and PCH Double loses this advantage. Indeed, the fraction of subjects that best respond in all games drops by 25% as we move from four games to twelve.

For PCH Single more games proves helpful. This is because the probability of perfectly best responding to all game is still decreasing, but now the probability is $(1-\epsilon)^g$. Thus, the probability decreases more slowly compared to the first case, so the effect is smaller. Also the over-fitting problem is less likely to happen. First, this can be due to a higher probability of deviations and second, due to the absence of τ, which allows PCH to have more diverse prediction. Thus, the over-fitting effect dominates the perfectly best response effect, which results in the opposite comparative static, compared to PCH Double.

For the second case (panel b), consider QRE versus LK Double when there are only four games. Of the subjects for whom QRE performs better, 73% have a $-\infty$ likelihood value for LK Double. This happens because the QRE DGP often plays Nash equilibrium in three of the four games, in which case LK Double is estimated to be a deterministic model ($\hat{\epsilon} = 0$). As the number of games grows, however, it becomes less likely that LK Double is so severely penalized, causing it to win (and QRE to lose) more frequently.

Figure II shows the same graphs when we apply BIC. Here a larger number of games is always helpful. This again is due to severeness of punishment for LOOCV. For BIC, as we increase the number of games, each model and QRE become more distinctive from each other, leading to improvements in model selection without being tainted by



(a) (b)

FIGURE II. Winning frequencies when using BIC. Panel (a): Winning frequencies of each model as the DGP. Panel (b): Winning frequency of QRE as the DGP against each model.

severe penalties. One thing to notice is that for any number of games, the performance is better when the DGP is QRE. This is because the BIC punishes based on the number of parameters and QRE has uniquely the smallest number of parameters.

Number of Strategies

Intuitively, games with a larger strategy space should allow for better model distinction. To test this, we switch from 3 × 3 games to two-person guessing games (Costa-Gomes and Crawford, 2006) that have an interval strategy space. We can then study the effect of the strategy space by placing a grid on the interval and varying only the coarseness of that grid.

Briefly, in a two-person guessing game each player i is given the interval $[a_i, b_i]$ and target p_i. They select a guess $s_i \in [a_i, b_i]$ and are paid based on how far their guess is from $p_i s_j$, which is their target times the other's guess. Closer guesses receive higher payments. Iterative logic is natural here, so levels-based models typically perform well (Costa-Gomes and Crawford, 2006), though there is evidence that parameters are not stable across games (Georganas et al., 2015). Further details about these games appear in the appendix.

We first repeat our simulation exercise on guessing games with a fine strategy space, meaning subjects can pick any integer in $[a_i, b_i]$. Simulated subjects play 12 such games that vary in their intervals and targets. The intervals vary in width from 200 to 800. Then we repeat the exercise with a coarse strategy space in which subjects must pick

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	HQR
LK double	**86.7** **83.4**					
LK single		**89.2** **84.0**				
PCH double			82.1 82.2			
PCH single				**81.8** **77.1**		
QLK					76.0 75.8	
HQR						72.5 71.9
QRE	**77.6** **71.0**	**85.0** **77.2**	94.3 95.2	94.4 94.3	94.7 94.9	93.6 94.3

TABLE XII. LOOCV winning frequency of each model versus QRE in the guessing games. Top number: Fine strategy space. Bottom number: coarse strategy space. Bottom row: QRE as the DGP versus each model. Bold: Significant difference with $p < 0.01$. Italics: Significant with $p \in [0.01, 0.05)$.

numbers that are multiples of ten. Thus, the coarse strategy spaces have between 21 and 81 strategies.

The winning frequencies of each model as the DGP versus QRE are shown in Table XVI, with QRE as the DGP in the bottom row. The top number in each cell represents the result with the fine strategy sets and the bottom number shows the results for coarse strategy sets. The bold numbers indicate that the difference between fine strategy sets and coarse strategy sets is significant at the 1% level according to a Fisher's exact test.[22]

One might guess that the games with fine strategy sets would provide significantly better model identification. While this is generally true, more than half of the differences (7 out of 12) are not significant. Thus, increasing the number of strategies often did not substantially improve model selection performance in our simulations.

Table XIII shows the results when we compare with BIC. Bold entries are significantly different at the 1% level and italicized entries are significant at the 5% level. Like with LOOCV, the fine strategy space significantly improves performance in 5 of the 12 comparisons. Unlike LOOCV, there are actually two cases where the fine strategy space performs significantly worse under BIC. Comparing across methods, however, the BIC does perform better than LOOCV for all models except QLK and HQR. This echos our finding that BIC generally performs better than LOOCV, but seems to benefit less from having a larger strategy space.

[22] All other results are not significant, even at the 10% level.

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	HQR
LK double	**98.3** **96.9**					
LK single		**98.2** **96.4**				
PCH double			**87.4** **92.5**			
PCH single				**87.1** **91.7**		
QLK					65.8 65.8	
HQR						69.3 69.4
QRE	*98.0* *97.0*	96.4 95.7	*99.7* *99.1*	**99.7** **99.0**	99.6 99.5	98.7 98.1

TABLE XIII. A replication of Table XVI with the BIC instead of LOOCV.

The fact that the BIC does not always perform better on larger strategy spaces may reflect the fact that the overfitting penalty ideally should adjust as the strategy space changes, but the BIC penalty does not. Results for the AIC appear in the appendix and are very similar to the BIC.

Lesson 6. The ability of models to overfit data can depend on the size of the strategy space, but the BIC and AIC overfitting penalties do not account for this.

VI. DISCUSSION

In this paper, we investigate whether model selection can correctly identify the behavioral game theory model that generated a given dataset. We find that ties between models can generate significant identification problems, leading to failures in model selection. In our case, simply rescaling the payoffs was not sufficient to avoid the problem. Instead, restricting the model selection exercise to only two competing models proved largely successful. There were, however, a handful of noticeable failures when using cross validation, even in the absence of ties. These failures were specific to the cross validation methods used and the models under consideration. Most often failures occurred when one model was infinitely penalized because it was estimated to be a deterministic model whose out-of-sample (pure strategy) prediction then proved wrong.

The infinite-penalty problem can be avoided mechanically by restricting the parameter grids when estimating the models. For example, requiring $\epsilon \geq 0.1$ and $\lambda \leq 10$ prevents the models from making deterministic out-of-sample predictions. We find this solution undesirable, however, because the parameter bounds would be *ad hoc* and may artificially penalize one model over another. In addition, many researchers are particularly interested in those subjects whose play perfectly fits a deterministic model. Much of the analysis of Costa-Gomes and Crawford (2006), for example, is focused on those players that always play according to a given level. Ruling out extreme parameters would misidentify these subjects. Finally, such parameter grid restrictions would only address the first anomaly, but not the second or third.

Another possible solution would be to incorporate regularization, in which model parameters are penalized for fitting "extreme observations." For example, a model's parameters could be penalized when the resulting model generates deterministic (or, nearly deterministic) predictions. Of course, which regularization penalty is added will depend on which anomalies one seeks to avoid in their model selection exercise; a necessary first step is therefore to identify the problematic anomalies that need to be avoided.

Given that the BIC and AIC outperform LOOCV in our simulations, perhaps one should eschew cross-validation methods in favor of the BIC or AIC. Although these do perform better in our 3×3 games when comparing only two models, they do not perform predictably as the size of the strategy space changes in guessing games. This suggests that the overfitting penalty ideally should adjust as the strategy space changes, but the BIC and AIC are inflexible and thus their performance can vary noticeably. A benefit of cross validation is that its overfitting "penalty" is endogenous: models are implicitly penalized for failing to predict ouf-of-sample, and this penalty naturally varies as the strategy space (and thus, the degree of overfitting) changes.

Given that cross validation and BIC/AIC methods both suffer weaknesses, our best recommendation is that researchers simulate the model selection exercise in their own setting to see which issues are potentially problematic. Only then can they know if their models face serious identification challenges, if cross validation suffers by excessively penalizing some models, and if the BIC and AIC perform appropriately for their environment.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE transactions on automatic control 19, 716–723.

Balietti, S., Klein, B., Riedl, C., 2021. Optimal design of experiments to identify latent behavioral types. Experimental Economics 24, 772–799. doi:10.1007/s10683-020-09680-w.

Breitmoser, Y., 2012. Strategic reasoning in p-beauty contests. Games and Economic Behavior 75, 555–569.

Camerer, C.F., Ho, T.H., Chong, J.K., 2004. A Cognitive Heirarchy Model of Games. Quarterly Journal of Economics 119, 861–898.

Carbone, E., 1997. Discriminating Between Preference Functionals: A Monte Carlo Study. Journal of Risk and Uncertainty 15, 29–54. doi:10.1023/A:1007785820094.

Carbone, E., Hey, J.D., 1994a. Discriminating between preference functionals: A preliminary monte carlo study. Journal of Risk and Uncertainty 8, 223–242.

Carbone, E., Hey, J.D., 1994b. Estimation of Expected Utility and Non-Expected Utility Preference Functionals Using Complete Ranking Data, in: Munier, B., Machina, M.J. (Eds.), Models and Experiments in Risk and Rationality. Springer Netherlands, Dordrecht. Theory and Decision Library, pp. 119–139.

Chen, C.T., Huang, C.Y., Wang, J.T.y., 2018. A window of cognition: Eyetracking the reasoning process in spatial beauty contest games. Games and Economic Behavior 111, 143–158.

Costa-Gomes, M., Crawford, V.P., 2006. Cognition and Behavior in Two-Person Guessing Games: An Experimental Study. American Economic Review 96, 1737–1768.

Costa-Gomes, M., Crawford, V.P., Broseta, B., 2001. Cognition and Behavior in Normal-Form Games: An Experimental Study. Econometrica 69, 1193–1235.

El-Gamal, M.A., Palfrey, T.R., 1996. Economical experiments: Bayesian efficient experimental design. International Journal of Game Theory 25, 495–517.

Feltovich, N., 2000. Reinforcement-based vs. belief-based learning models in experimental asymmetric-information games. Econometrica 68, 605–641.

Fudenberg, D., Gao, W., Liang, A., 2020. How Flexible is that Functional Form? Quantifying the Restrictiveness of Theories. arXiv preprint arXiv:2007.09213 `arXiv:2007.09213`.

Fudenberg, D., Kleinberg, J., Liang, A., Mullainathan, S., 2022. Measuring the completeness of economic models. Journal of Political Economy 130, 956–990.

García-Pola, B., Iriberri, N., Kovářík, J., 2020. Non-equilibrium play in centipede games. Games and Economic Behavior 120, 391–433. doi:10.1016/j.geb.2020.01.007.

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

Goeree, J.K., Holt, C.A., 2004. A model of noisy introspection. Games and Economic Behavior 46, 365–382.

Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307. doi:10.1093/biomet/76.2.297.

Jakusch, S.T., 2013. On the Applicability of Maximum Likelihood Methods: From Experimental to Financial Data. doi:10.2139/ssrn.2845871.

Johnson, E.J., Camerer, C., Sen, S., Rymon, T., 2002. Detecting failures of backward induction: Monitoring information search in sequential bargaining. Journal of Economic Theory 104, 16–47.

McKelvey, R.D., Palfrey, T.R., 1995. Quantal Response Equilibria for Normal Form Games. Games and Economic Behavior 10, 6–38.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. John Wiley & Sons, New York.

Nagel, R.C., 1995. Unraveling in Guessing Games: An Experimental Study. American Economic Review 85, 1313–1326.

Rao, R.B., Fung, G., Rosales, R., 2008. On the dangers of cross-validation. An experimental evaluation, in: Proceedings of the 2008 SIAM International Conference on Data Mining, SIAM. pp. 588–596.

Salmon, T.C., 2001. An evaluation of econometric models of adaptive learning. Econometrica 69, 1597–1628.

Schwarz, G., 1978. Estimating the Dimension of a Model. The Annals of Statistics 6, 461–464. doi:10.1214/aos/1176344136.

Shao, J., 1997. An asymptotic theory for linear model selection. Statistica sinica , 221–242.

Stahl, D.O., Wilson, P.O., 1994. Experimental Evidence on Players' Models of Other Players. Journal of Economic Behavior and Organization 25, 309–327.

Stahl, D.O., Wilson, P.W., 1995. On Players' Models of Other Players: Theory and Experimental Evidence. Games and Economic Behavior 10, 218–254.

Stone, M., 1977. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. Journal of the Royal Statistical Society: Series B (Methodological) 39, 44–47. doi:10.1111/j.2517-6161.1977.tb01603.x.

Turocy, T.L., 2005. A dynamic homotopy interpretation of the logistic quantal response equilibrium correspondence. Games and Economic Behavior 51, 243–263. doi:10.1016/j.geb.2004.04.003.

Wright, J.R., Leyton-Brown, K., 2017. Predicting human behavior in unrepeated, simultaneous-move games. Games and Economic Behavior 106, 16–37.

Yang, Y., 2005. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. Biometrika 92, 937–950.

Appendix A. Guessing Game Structures

Here we describe the two-person guessing games from Georganas et al. (2015) and Costa-Gomes and Crawford (2006). Two-person guessing games are described as follows. Each player i is given the interval $[a_i, b_i]$ and target p_i. $s_i \in [a_i, b_i]$ denotes a strategy or a guess and payment is determined by how far i's guess from i's target times j's guess. Specifically, let $e_i = |s_i - p_i s_j|$. Then the payoff function is,

$$u_i(e_i) = \begin{cases} 15 - \frac{11}{200}e_i & \text{if } e_i \le 200 \\ 5 - \frac{1}{200}e_i & \text{if } e_i \in [200, 1000) \\ 0 & \text{otherwise} \end{cases}$$

We use the following intervals and targets for each game.

Table XIV. Two-Person Guessing Games

	Player's Limits & Targets	Opponent's Limits & Targets
Game 1	([100,500], 0.7)	([100,900], 1.3)
Game 2	([100,500], 0.5)	([300,900], 1.3)
Game 3	([300,500], 1.5)	([100,500], 1.5)
Game 4	([100,500], 0.7)	([100,500], 1.5)
Game 5	([300,500], 0.7)	([100,900], 1.3)
Game 6	([300,900],1.3)	([100,900], 1.5)
Game 7	([100,900],1.3)	([100,500], 0.7)
Game 8	([300,900], 1.3)	([100,500], 0.5)
Game 9	([100,500], 1.5)	([300,500], 1.5)
Game 10	([100,500], 1.5)	([100,500], 0.7)
Game 11	([100,900], 1.3)	([300,500], 0.7)
Game 12	([100,900], 1.5)	([100,900], 1.3)

Appendix B. Additional Results

Expected Discrimination of Games

Figures III and IV show the distributions of the expected discrimination for each pair of models across the randomly-chosen games. The vertical green line shows the 90th percentile of each distribution. Stars indicate the expected discrimination of our original 12 games; game numbers appear above each star. For each pair of models there are multiple stars above the green dashed line, indicating that the original 12 games provided reasonable expected discrimination.

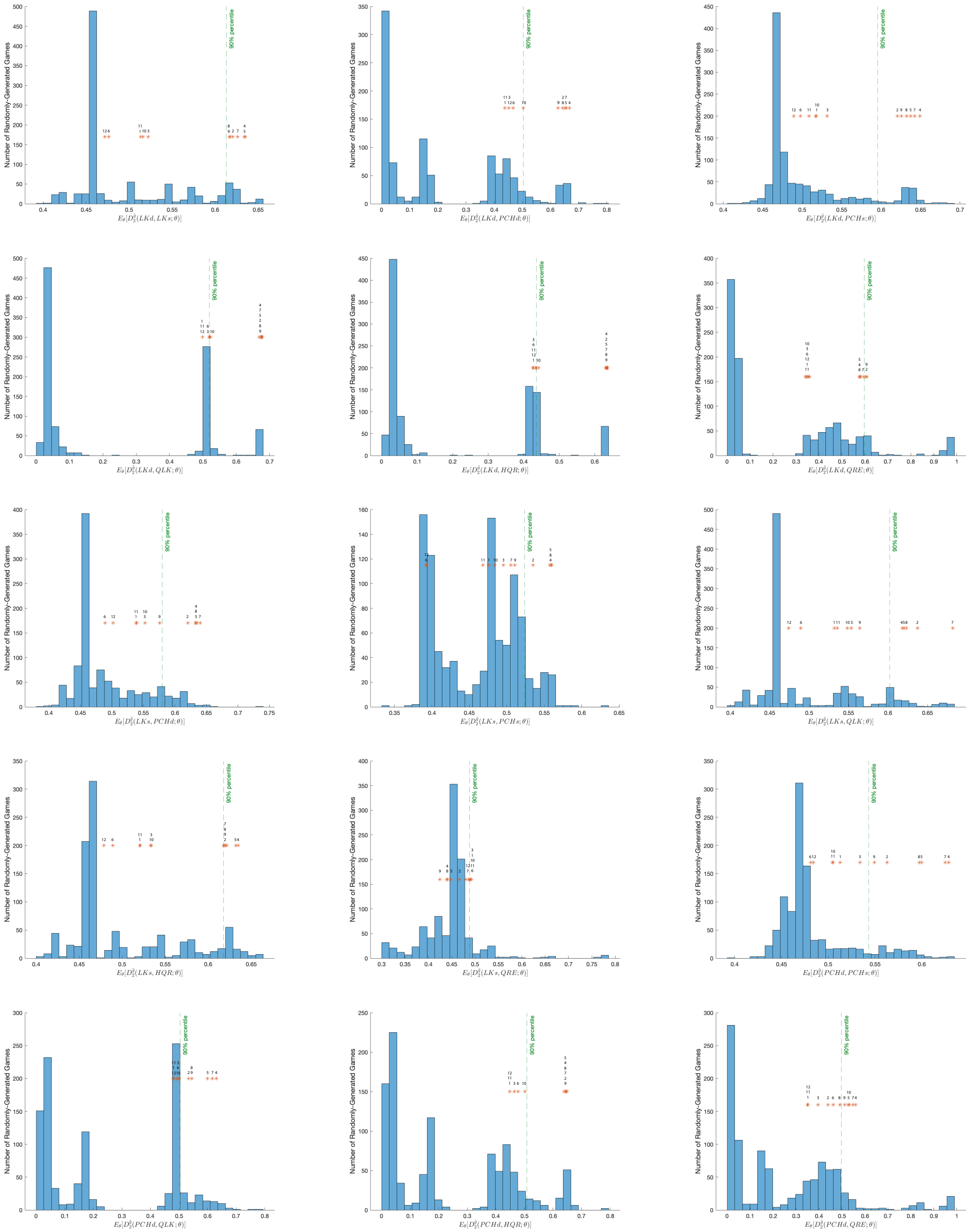


FIGURE III. Distributions of $E_\theta[D_2^{\tilde g}(M_k,M_l;\theta_k,\theta_l)]$for each pair of models (M_k,M_l).

Number of Games

For both LOOCV and BIC comparisons, trends are quite similar to original payoffs but the absolute values for each winning probabilities are much smaller.

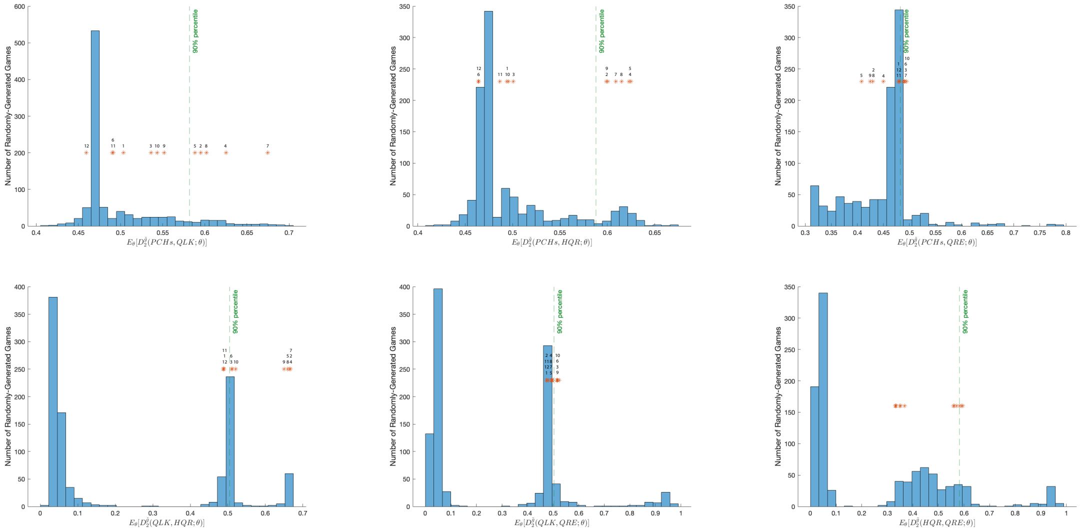
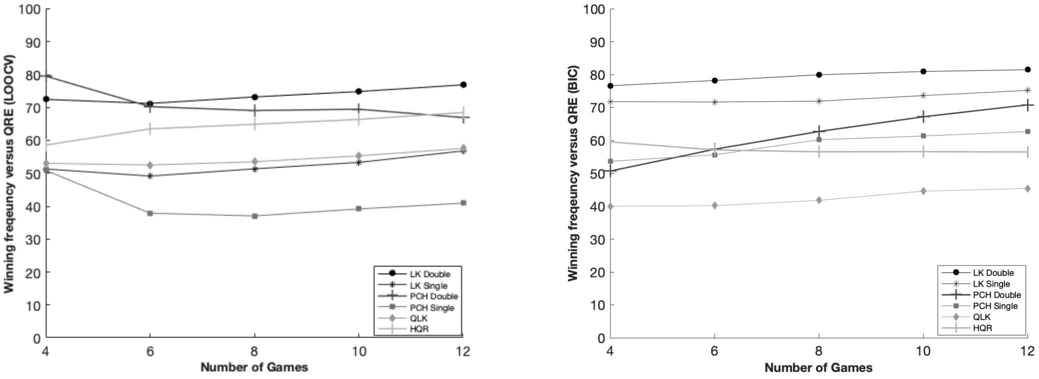


FIGURE IV. Distributions of $E_\theta[D_2^{\tilde{g}}(M_k, M_l; \theta_k, \theta_l)]$for each pair of models (M_k, M_l).

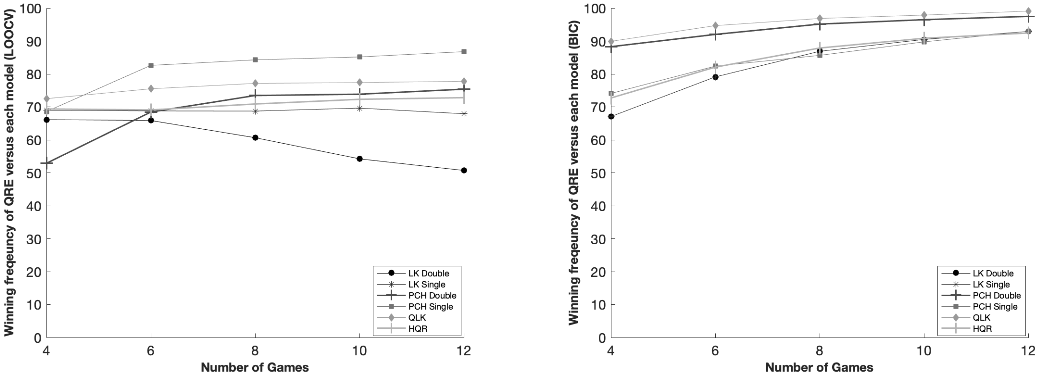


(a)

(b)

FIGURE V. Winning frequencies of each of the levels-based models vs. QRE as the number of games varies when we scale the payoff by 1/00. Panel (a): LOOCV. Panel (b): BIC.

This is related to how logistic errors/responses work. For the models with deterministic responses, it is more likely to deviate for double counting models and even deviate more to the worst strategies for single counting models. For the models with non-deterministic responses also show more “random" movements including QRE itself. (Especially recall that HQR and QLK can behave like LK models without errors even for moderate λ) Thus, QRE and other models now behave more similarly which increases the probability of wrong model wins.

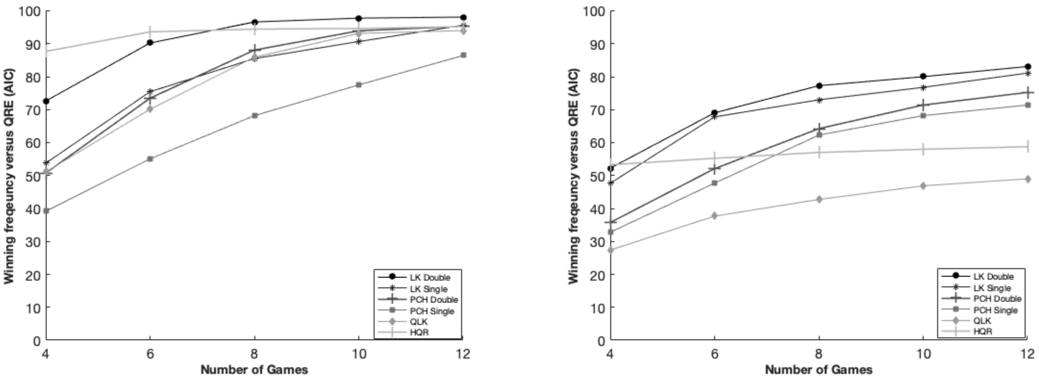


(a) (b)

FIGURE VI. Winning frequencies of QRE vs. each of the levels-based models as the number of games varies when we scale the payoff by 1/00. Panel (a): LOOCV. Panel (b): BIC.

AIC Results

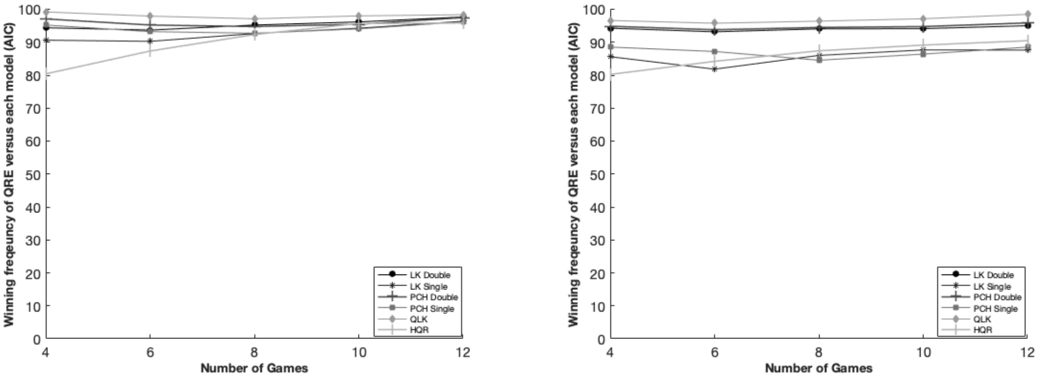
AIC shows quite similar results with BIC. Increasing number of games can be helpful, and scaling the payoff by 1/100 can make the results worse.



(a)

(b)

FIGURE VII. Winning frequencies of each of the levels-based models vs. QRE as the number of games varies when we use AIC Panel (a): Original payoff. Panel (b): payoffs are scaled by 1/100.



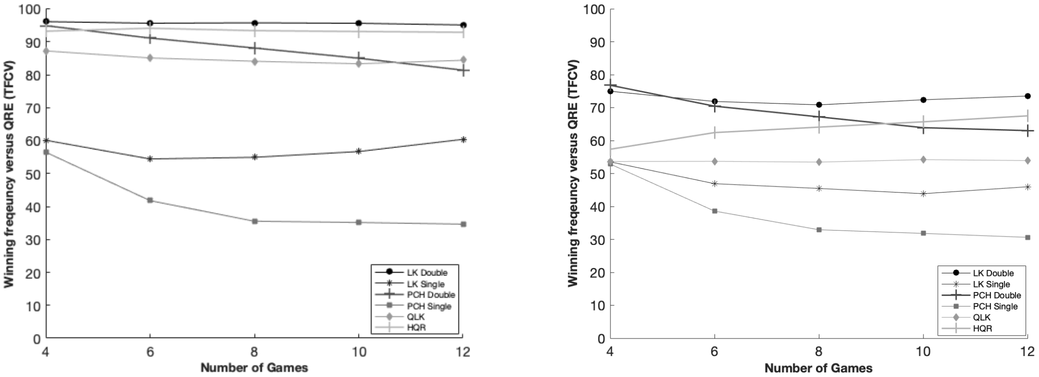
(a)

(b)

FIGURE VIII. Winning frequencies of QRE vs. each of the levels-based models as the number of games varies when we use AIC Panel (a): Original payoff. Panel (b): payoffs are scaled by 1/100.

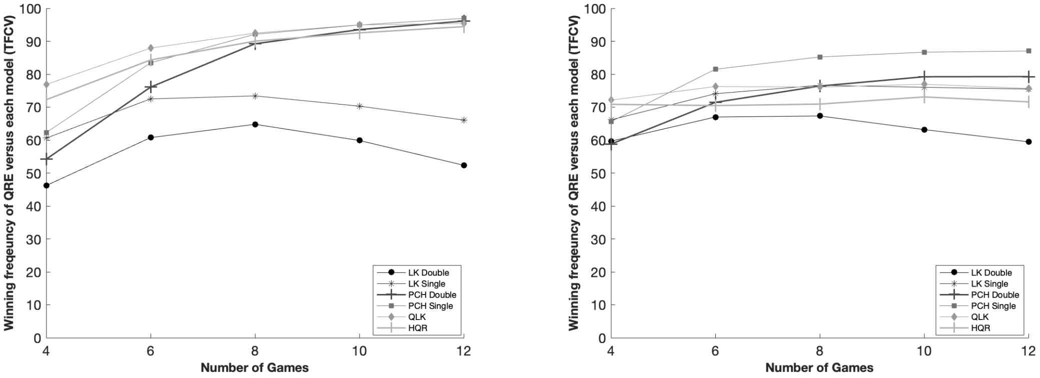
2FCV Results

For two-fold cross validation, the benefits of both increasing number of games and making payoff smaller are not very clear. Two-fold cross validations have similar structure with LOOCV, but less severe punishment compare to LOOCV.



(a) (b)

FIGURE IX. Winning frequencies of each of the levels-based models vs. QRE as the number of games varies when we use 2FCV Panel (a): Original payoff. Panel (b): payoffs are scaled by 1/100.



(a) (b)

FIGURE X. Winning frequencies of QRE vs. each of the levels-based models as the number of games varies when we use 2FCV Panel (a): Original payoff. Panel (b): payoffs are scaled by 1/100.

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	HQR
LK double	**78.63** **76.43**					
LK single		**46.87** **79.93**				
PCH double			*50.1* *54.23*			
PCH single				*71.87* *78.93*		
QLK					**66.6** **66.16**	
HQR						**67.5** **66.53**
QRE	**76.67** **69.7**	**98.9** **60.73**	*96* 95.37	*93.93* *95.33*	93 92.77	93.43 92.5

TABLE XV. 2FCV winning frequency of each model versus QRE in the guessing games. Top number: Fine strategy space. Bottom number: coarse strategy space. Bottom row: QRE as the DGP versus each model. Bold: Significant difference with $p < 0.01$. Italics: Significant with $p \in [0.01, 0.05)$.

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	HQR
LK double	**98.67** **97.5**					
LK single		**98.57** **97.07**				
PCH double			**88.1** **93.73**			
PCH single				**87.9** **93.57**		
QLK					68.67 68.73	
HQR						72.17 72.1
QRE	96.57 97.13	94.5 95.37	99.17 98.93	99.07 98.83	99.27 99.2	98 98.27

TABLE XVI. AIC winning frequency of each model versus QRE in the guessing games. Top number: Fine strategy space. Bottom number: coarse strategy space. Bottom row: QRE as the DGP versus each model. Bold: Significant difference with $p < 0.01$. Italics: Significant with $p \in [0.01, 0.05)$.

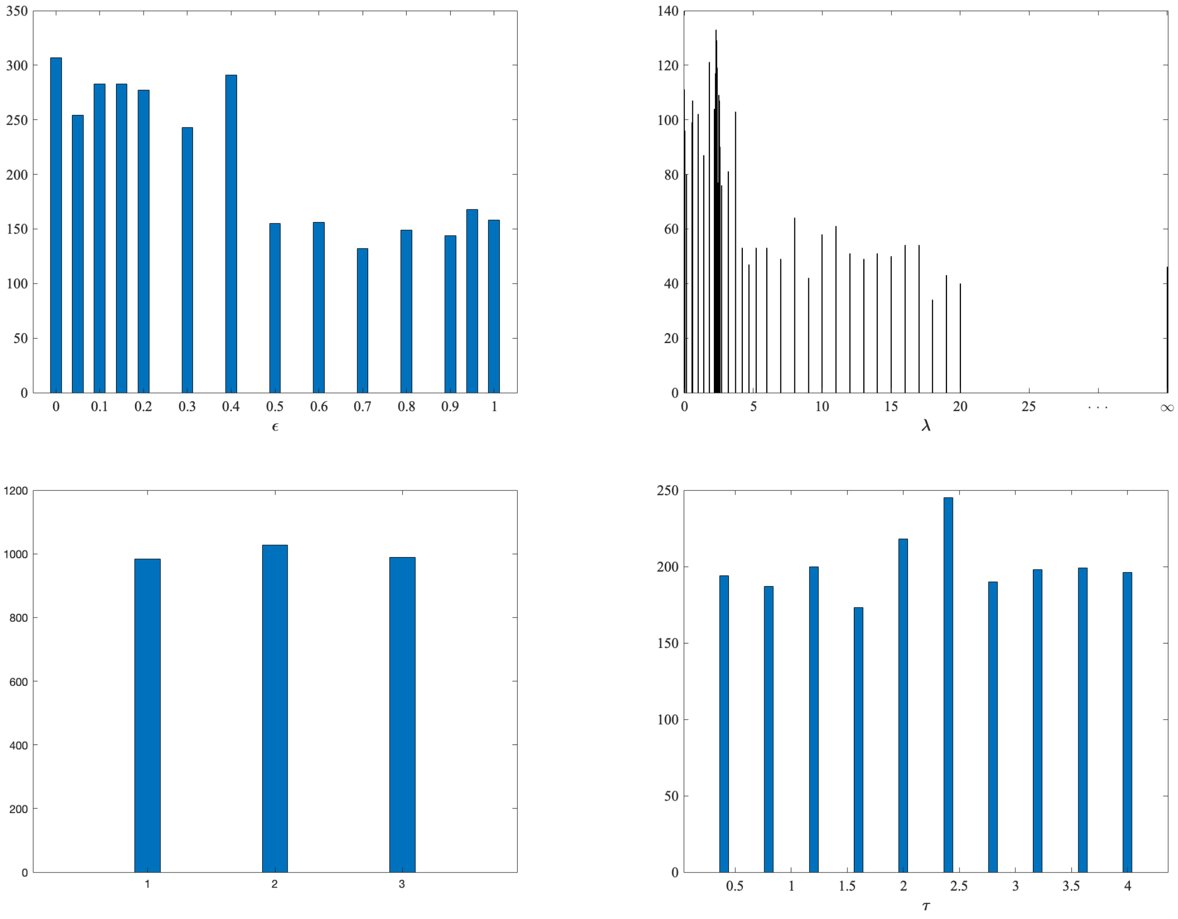


FIGURE XI. Frequency of true parameter values in PCH Single

APPENDIX C. NON-UNIFORM AND UNIFORM PARAMETER DISTRIBUTIONS

For the levels-based models our parameter distributions were not quite uniform over the parameter grids. This is because we drew parameters for simulated subjects using a three step procedure. First, we generated 2,000 level-1 subjects using truly uniform and independent draws over the other parameter grids. Then we copied those drawn parameter values for level 2 and level 3 (if applicable). This led to 6,000 total subjects and guaranteed that the distribution of λ values was exactly the same for level 1, level 2, and level 3. Finally, we randomly drew 3,000 subjects out of this pool of 6,000 for our final sample. This process can deviate from a true uniform distribution, however, because sampling error in the initial draw of 2,000 subjects gets amplified by using three copies of that sample. In our case, lower values of ϵ and λ happened to be drawn more frequently. Similarly, any slight correlation between parameters due to sampling error also gets amplified in the final sample. The marginal distributions of the parameters for

	k	ϵ	λ	τ
LK Double	0.7225	0.1018	0.7545	
LK Single	0.4912	0.5618	0.0545	
PCH Double	0.3779	0.8031	0.1794	0.4101
PCH Single	0.6845	0.3618	0.5735	0.3798

QLK	0.1251		0.9115 (λ^1)
			0.4687(λ^2)
			0.1788($\lambda^{1(2)}$)	

HQR	0.2483	0.3684		
QRE			0.6219	

TABLE XVII. p-values for chi-squared goodness of fit tests that each parameter distribution is drawn uniformly.

PCH Single (for example) are shown in Figure XI. The correlation between λ and ϵ is 0.0567, which is significant with a p-value of 0.0019.

We can use this fact to explore the sensitivity of our main results to changes in the distribution. To that end, we re-run the main part of our exercise using 3,000 new simulated subjects for each model, where the parameters are now drawn from truly uniform and independent distributions. In this case the level 2 subjects' parameters are drawn independently of the level 1 subjects' parameters. To ensure our process is truly uniform we verify in Table XVII that the new distributions are not significantly different from uniform using χ^2 goodness-of-fit tests.

Now we ask whether our main results change with truly uniform distributions. Table XVIII replicates Table IV. Still we see a large problem with similar models becoming indistinguishable. And Table XIX verifies that switching to small payoffs does not solve the problem.

Table XX shows that, when comparing only two models, the same anomalies occur. With uniformly-drawn parameters PCH Single now wins just over 50% when compared to QRE, but we still consider it an anomaly since its success rate is no different from a coin flip.

Finally, the BIC success rates are shown in Table XXI. Again, this method avoids the anomalies caused by cross validation. Thus, our main conclusions appear not to be overly sensitive to the distribution of simulated subjects' parameters.

DGP	win type	LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
LK Double	solo	0.7	0.63	0.5	0.27	0.37	0.47	2.47
	tie-DGP	N/A	92.87	62.17	62.13	62.17	93.23	0.2
	tie-others	93.3	0.13	0.9	1	0.3	0.3	0
LK Single	solo	4.27	42.93	3.67	3.97	3.27	5.6	8.03
	tie-DGP	17.6	N/A	11.87	18.17	11.87	17.57	1.57
	tie-others	1.93	23.9	1.73	0.37	2.5	2.37	0
PCH Double	solo	2.23	1.3	1	1.2	7.3	2.23	4.43
	tie-DGP	46.03	45.77	N/A	78.73	46.4	45.77	1.1
	tie-others	0.83	0	79	0	0.47	1.3	0
PCH Single	solo	8.47	19.2	2.4	16.87	5.53	8.07	10.5
	tie-DGP	7.73	14.8	17.07	N/A	8.23	7.73	0.67
	tie-others	2.73	0.1	1.63	24.13	2.6	2.83	0
QLK	solo	0.87	1.9	0.37	0.67	1.63	1.27	4.57
	tie-DGP	86.03	85.93	87.23	87.17	N/A	86.33	0.4
	tie-others	1	0.07	0.27	0.13	87.6	0.8	0
HQR	solo	1.17	0.7	0.6	0.67	0.5	2.4	3.13
	tie-DGP	88.3	87.77	58.77	58.63	59.27	N/A	0.03
	tie-others	0.13	0.1	1.93	1.9	0	88.8	0
QRE	solo	42.4	0.83	0.33	0.67	1.03	2.1	21.93
	tie-DGP	0	0	0	0	0	0	N/A
	tie-others	29.7	29.73	0.9	1.03	0.7	0.7	0

TABLE XVIII. For each DGP (row) and model (column), the percentage of subjects for which the model wins uniquely ("solo"), wins in a tie with the DGP (and possibly other models; "tie-DGP"), and wins in a tie with other non-DGP models ("tie-others"). Parameters are drawn from a true uniform distribution.

DGP	win type	LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
LK Double	solo	7.73	6.6	4.33	2.37	3.13	4.83	10.9
	tie-DGP	N/A	54.6	38.13	36.9	37.23	55.13	1.83
	tie-others	56.6	0.57	1.5	2.03	1.43	1.47	0
LK Single	solo	8.3	28.87	6.2	5.03	4.13	8.2	12.4
	tie-DGP	17.6	N/A	11.9	16	11.9	17.6	1.57
	tie-others	2.9	21.7	2.97	0.33	2.27	2.27	0
PCH Double	solo	8.87	5.7	3.4	4.4	5.67	6.97	13.4
	tie-DGP	27.63	25.77	N/A	46.97	25.87	25.87	1.2
	tie-others	0.8	0.93	48.87	0.8	1.23	1.87	0
PCH Single	solo	10.1	14.37	4.77	11.83	6.43	9.3	15.17
	tie-DGP	7.67	12.27	16.9	N/A	7.67	7.67	0.6
	tie-others	4.1	0.2	3.2	21.5	2.93	3.2	0
QLK	solo	8.2	7.17	3.2	4.73	6.03	14.47	17
	tie-DGP	26.07	25.47	25.7	25.47	N/A	29.03	0.27
	tie-others	4.33	0.8	6.17	5.63	29.23	3.07	0
HQR	solo	8.07	6.13	3.13	4.83	4.77	14.47	16.03
	tie-DGP	37.1	35.53	23.93	23.73	26.43	N/A	0.07
	tie-others	2	0.53	2.33	1.1	0.23	39.47	0
QRE	solo	27.03	6.33	4.17	5	5	10.83	25.23
	tie-DGP	0.03	0.03	0.03	0.03	0.03	0.03	N/A
	tie-others	13.83	12.4	1.57	0.83	2.37	2.5	0.03

TABLE XIX. A replication of Table IV when game payoffs are scaled by 1/100.

DGP\EST	LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
LK Double	95.25						4.75
LK Single		75.63					24.37
PCH Double			82.18				17.82
PCH Single				*50.7*			49.3
QLK					91.55		8.45
HQR						92.38	7.62
QRE QRE wins:	*26.27*	*50.37*	97.03	96.98	96.13	94.53	
other wins:	73.73	49.63	2.97	3.02	3.87	5.47	

TABLE XX. LOOCV winning frequency of each model versus only QRE in the 3 ×3 games with the original payoffs.

DGP\EST	LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
LK Double	98						2.47
LK Single		91.77					8.23
PCH Double			94.4				5.6
PCH Single				73.03			26.97
QLK					94.7		5.3
HQR						95	5
QRE QRE wins:	97.87	96.33	98.3	97.73	98.6	97	
Model wins:	2.13	3.67	1.7	2.27	1.4	3	

TABLE XXI. BIC winning frequency of each model versus only QRE in the 3 ×3 games with the original payoffs

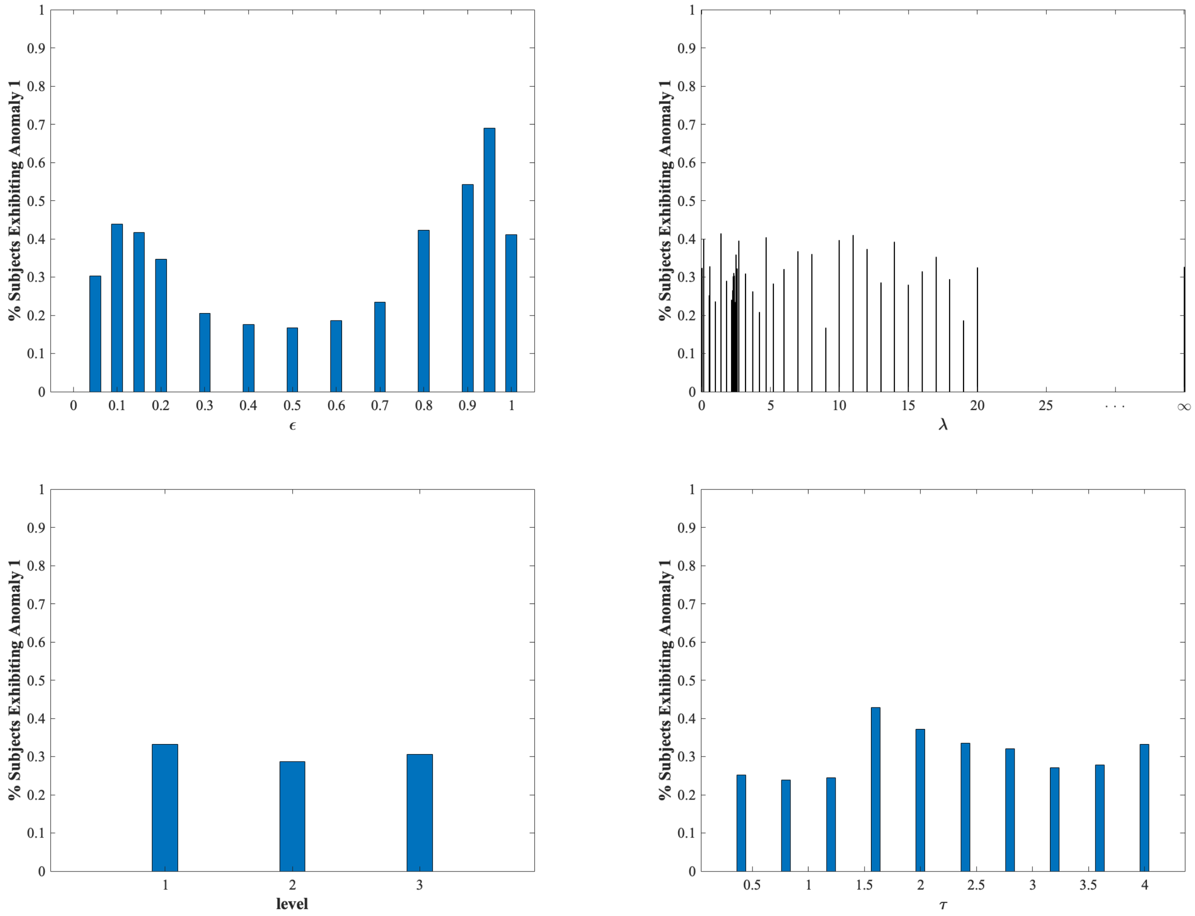


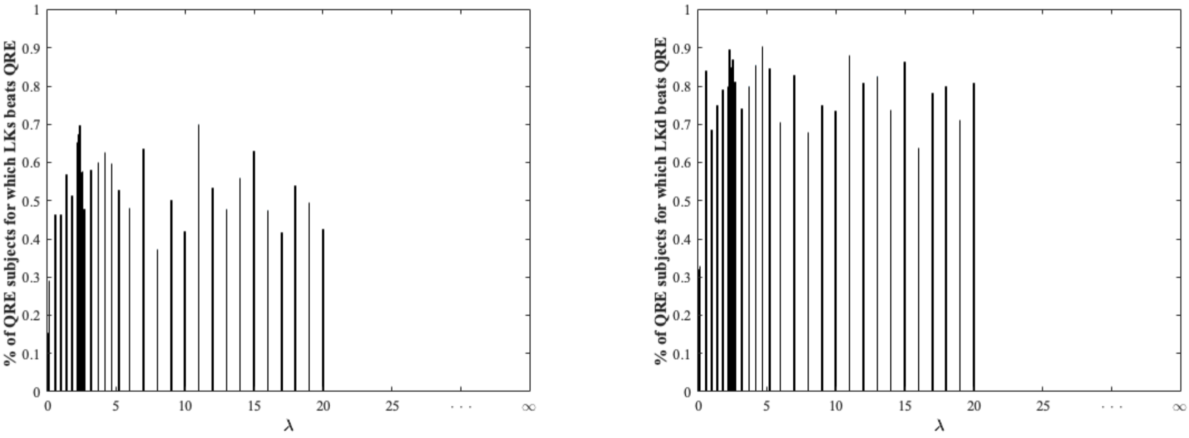
FIGURE XII. For each true parameter value in PCH Single, the fraction of subjects at that parameter value who exhibit Anomaly #1.

Online Appendix

APPENDIX D. PARAMETER VALUES THAT GENERATE ANOMALIES

Figure XII shows how frequently a PCH Single subject exhibits Anomaly #1 (infinite likelihood penalties), broken down by the subject's true parameter values.

Anomalies 2 and 3 are harder to measure directly since they impact single folds of the cross-validation procedure, and it's hard to quantify the impact of any single fold on the overall result. For simplicity, we present instead the frequency with which the wrong model wins for each true λ with the QRE DGP. This is shown in Figure XIII. The left panel shows how frequently LK Single beats QRE when QRE is the DGP. The vast majority of these observations are due to Anomaly 2, and Anomaly 3 is excluded since it only applies to LK Double. The right panel considers QRE versus LK Double and therefore includes both Anomaly 2 and Anomaly 3. In both figures the frequency



QRE vs. LK Single

QRE vs. LK Double

FIGURE XIII. For each true λ in QRE, the fraction of subjects for which the wrong model wins.

is fairly constant in λ, except for the lowest values at which equilibrium play becomes quite rare.

APPENDIX E. EXTRA RESULTS AND TABLES

Table XXII, Table XXIII, Table XXIV report the winning frequency of each model when we control for ties, and when we use 2FCV, BIC, or AIC.

The winning frequency of each DGP model without ties is quite low in almost all cases.

DGP	win type	LK Double	LK Single	PCH Double	PCH Single	QLK	QR	QRE
LK Double	solo	0.83	0.47	0.37	0.23	0.47	0.57	2.73
	tie-DGP	N/A	92.7	62.2	50.6	45.57	92.9	0.43
	tie-others	93.1	0.07	1.1	1	0.23	0.07	0
LK Single	solo	4.63	35.2	4.73	2.87	5.87	5.3	17.67
	tie-DGP	16.2	N/A	11.67	14.47	9.47	16.17	2.13
	tie-others	1.97	20.33	2.17	0.5	1.5	1.43	0
PCH Double	solo	2.33	1	8.23	0.6	7.17	1.9	6.9
	tie-DGP	46.6	46.43	N/A	65.27	39.47	46.43	1.3
	tie-others	0.47	0.13	71.07	0.13	0.2	0.67	0
PCH Single	solo	7.87	14.83	6.33	10.43	9.37	7	21.43
	tie-DGP	8.1	11.2	14.4	N/A	7.7	8.1	1.6
	tie-others	3	0.53	2.53	17.5	2.73	2.93	0
QLK	solo	1.27	0.67	0.6	0.37	2.27	1.27	5.27
	tie-DGP	62.7	62.63	63.5	55.27	N/A	63	0.27
	tie-others	23.77	23.33	24	14.9	63.83	23.67	0
QR	solo	0.63	0.63	0.43	0.37	0.87	1.5	3.93
	tie-DGP	89.2	89.07	60.2	49	43.87	N/A	0.27
	tie-others	0.17	0.03	2	1.53	0.37	89.6	0
QRE	solo	25.3	3.6	0.73	0.97	1.63	1.97	44.5
	tie-DGP	0.07	0.07	0.07	0.07	0.07	0.07	N/A
	tie-others	20.27	20.2	0.83	0.67	0.43	0.53	0.07

TABLE XXII. For each DGP (row) and model (column), the percentage of subjects for which the model wins uniquely (“solo”), wins in a tie with the DGP (and possibly other models; “tie-DGP”), and wins in a tie with other non-DGP models (“tie-others”) when using TFCV.

DGP	win type	LK Double	LK Single	PCH Double	PCH Single	QLK	QR	QRE
LK Double	solo	0.7	2.1	0	0.47	0.4	95.07	1.17
	tie-DGP	N/A	0	0	0	0.03	0	0
	tie-others	0.03	0	0.07	0.07	0	0	0
LK Single	solo	1.1	57.6	0.37	4.87	0.73	31.1	3.97
	tie-DGP	0	N/A	0	0	0	0	0
	tie-others	0.17	0	0.1	0.1	0.17	0	0
PCH Double	solo	0.4	1.9	12.13	4.5	1.83	49.57	3.9
	tie-DGP	0	0	N/A	25.77	0	0	0
	tie-others	0	0	25.77	0	0	0	0
PCH Single	solo	0.97	21.83	3.17	31.33	1.57	21.07	11.5
	tie-DGP	0	0	8.47	N/A	0	0	0
	tie-others	0.1	0	0	8.47	0.1	0	0
QLK	solo	0.17	1.33	0.1	0.73	4.8	89.03	3.7
	tie-DGP	0.1	0	0	0	N/A	0	0
	tie-others	0	0	0.03	0.03	0.1	0	0
QR	solo	0.1	0.93	0	0.97	0.3	92.5	5.17
	tie-DGP	0	0	0	0	0	N/A	0
	tie-others	0.03	0	0	0	0.03	0	0
QRE	solo	0.27	1.43	0.17	0.73	0.37	1.9	94.33
	tie-DGP	0	0	0	0	0	0	N/A
	tie-others	0	0	0.8	0.8	0	0	0

TABLE XXIII. For each DGP (row) and model (column), the percentage of subjects for which the model wins uniquely (“solo”), wins in a tie with the DGP (and possibly other models; “tie-DGP”), and wins in a tie with other non-DGP models (“tie-others”) when using BIC.

DGP	win type	LK Double	LK Single	PCH Double	PCH Single	QLK	QR	QRE
LK Double	solo	0.93	2.3	0	0.67	0.5	95	0.53
	tie-DGP	N/A	0	0	0	0	0	0
	tie-others	0	0	0.07	0.07	0	0	0
LK Single	solo	2.33	59.1	0.47	5.97	1.23	28.93	1.77
	tie-DGP	0	N/A	0	0	0	0	0
	tie-others	0.1	0	0.1	0.1	0.1	0	0
PCH Double	solo	0.6	2	14.07	5.27	1.8	47.8	2.7
	tie-DGP	0	0	N/A	25.77	0	0	0
	tie-others	0	0	25.77	0	0	0	0
PCH Single	solo	2.1	22.73	3.77	35.97	1.7	19.2	5.9
	tie-DGP	0	0	8.47	N/A	0	0	0
	tie-others	0.17	0	0	8.47	0.17	0	0
QLK	solo	0.3	1.73	0.27	1.23	5.1	89.2	2.07
	tie-DGP	0.07	0	0	0	N/A	0	0
	tie-others	0	0	0.03	0.03	0.07	0	0
QR	solo	0.2	1.4	0.03	1.27	0.7	92.87	3.53
	tie-DGP	0	0	0	0	0	N/A	0
	tie-others	0	0	0	0	0	0	0
QRE	solo	0.73	2.03	0.4	1.27	0.37	2.27	92.13
	tie-DGP	0	0	0	0	0	0	N/A
	tie-others	0	0	0.8	0.8	0	0	0

TABLE XXIV. For each DGP (row) and model (column), the percentage of subjects for which the model wins uniquely (“solo”), wins in a tie with the DGP (and possibly other models; “tie-DGP”), and wins in a tie with other non-DGP models (“tie-others”) when using AIC.

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	QR	QRE
LK double	21.48	21.35	11.14	11.12	11.05	21.18	2.67
LK single	9.34	50.05	6.26	7.71	6.29	9.89	10.46
PCH double	11.69	9.41	22.06	22.78	15.92	13.35	4.79
PCH single	9.16	22.51	10.42	22.17	9.78	11.7	14.25
QLK	15.6	15.98	14.99	15.06	16.44	15.67	6.26
QR	20.2	19.41	11	10.88	10.72	21.88	5.92
QRE	58.14	16.56	1.01	1.21	1.91	2.04	19.13

TABLE XXV. 3×3 original payoff LOOCV winning probabilities.

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	QR	QRE
LK Double	22.62	22.09	12.37	9.49	8.34	22.29	2.8
LK Single	9.15	40.81	7.47	6.9	8.16	9.54	17.97
PCH Double	10.83	9.25	28.71	19.71	13.99	10.42	7.09
PCH Single	10.56	17.85	11.93	16.45	11.9	9.66	21.66
QLK	17.34	16.5	17.13	12.82	13.49	17.43	5.3
HQR	21.46	21.34	12.49	9.64	8.65	22.45	3.97
QRE	35.39	13.66	1.11	1.3	1.83	2.19	44.51

TABLE XXVI. 3×3 original payoff 2FCV winning probabilities (including ties).

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	QR	QRE
LK double	0.93	2.3	0.03	0.7	0.5	95	0.53
LK single	2.38	59.1	0.52	6.02	1.28	28.93	1.77
PCH double	0.6	2	26.95	18.15	1.8	47.8	2.7
PCH single	2.18	22.73	8	40.17	1.78	19.2	5.93
QLK	0.33	1.73	0.28	1.25	5.13	89.2	2.07
QR	0.2	1.4	0.03	1.27	0.7	92.87	3.53
QRE	0.73	2.03	0.8	1.67	0.37	2.27	92.13

TABLE XXVII. 3×3 original payoff AIC winning probabilities.

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	QR	QRE
LK double	0.72	2.1	0.03	0.5	0.42	95.07	1.17
LK single	1.18	57.6	0.42	4.92	0.82	31.1	3.97
PCH double	0.4	1.9	25.02	17.38	1.83	49.57	3.9
PCH single	1.02	21.83	7.4	35.57	1.62	21.07	11.5
QLK	0.22	1.33	0.12	0.75	4.85	89.03	3.7
QR	0.12	0.93	0	0.97	0.32	92.5	5.17
QRE	0.27	1.43	0.57	1.13	0.37	1.9	94.33

TABLE XXVIII. 3×3 original payoff BIC winning probabilities.

DGP	# Parameters	LOOCV	2FCV	BIC	AIC
LK Double	3	8.07	7.2	7.37	7.8
LK Single	3	27.37	20.13	36.77	39.47
PCH Double	4	4.87	8.97	8.27	9.17
PCH Single	4	11.9	6.7	23.4	28.23
QLK	4	5.87	9	17.23	17.77
HQR	2	14.93	14.6	40.03	52.37
QRE	1	25.63	33.93	82.57	65.53

TABLE XXIX. Frequency with which each data generating process wins, excluding ties when payoffs are scaled by 1/100

DGP	# Parameters	LOOCV	2FCV	BIC	AIC
LK Double	3	56.3	56.03	0.33	0.43
LK Single	3	20.43	19.1	0	0
PCH Double	4	47.33	43.83	15.43	15.43
PCH Single	4	19.73	16.73	8.27	8.27
QLK	4	28.33	24	0.2	0.27
HQR	2	41.1	39.47	0	0
QRE	1	0	0.5	0	0

TABLE XXX. Frequency with which each data generating process wins in a tie with at least one other model. when payoffs are scaled by 1/100

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	QR	QRE
LK double	20.88	18.7	11.5	9.29	10.3	16.73	12.59
LK single	15.44	28.95	8.8	9.57	7.94	12.64	16.66
PCH double	13.16	9.87	18.79	16.91	11.93	12.78	16.56
PCH single	14.11	12.47	13.09	15.72	10.5	14	20.1
QLK	14.57	10.16	10.25	11.49	10.72	22.16	20.66
QR	17.86	14.63	7.12	8.63	8.93	24.66	18.18
QRE	33.38	12.09	4.58	6.51	6.75	13.15	23.53

TABLE XXXI. 3×3 payoffs are scaled by 1/100 LOOCV winning probabilities.

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	QR	QRE
LK Double	20.54	16.18	13.38	7.98	9.58	17.57	14.78
LK Single	14.13	25.09	10.13	6.99	8.97	11.76	22.93
PCH Double	13.09	7.89	22.28	14.63	12.78	11.94	17.38
PCH Single	13.09	13.1	14.09	12.17	10.79	11.97	24.8
QLK	15.19	9.6	11.18	8.88	13.88	17.67	23.61
HQR	17.12	13.24	9.37	6.41	9.97	24.39	19.5
QRE	23.13	8.64	5.6	4.3	7.89	12.6	37.86

TABLE XXXII. 3×3 payoffs are scaled by 1/100 2FCV winning probabilities.

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	QR	QRE
LK double	7.53	15.5	1.07	3.97	2	57.23	12.7
LK single	4.85	36.77	0.53	7.97	1.35	30.13	18.4
PCH double	2.62	8.4	15.98	17.18	3.02	33.67	19.13
PCH single	2.23	16.7	7.13	27.53	2.13	21.4	22.87
QLK	1.07	6.73	0.95	5.85	17.33	29	39.07
QR	0.9	7.1	0.47	5.63	1.03	40.03	44.83
QRE	0.87	5.27	0.73	3.6	1.93	5.03	82.57

TABLE XXXIII. 3×3 payoffs are scaled by 1/100 BIC winning probabilities.

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	QR	QRE
LK double	8.02	16.87	1.5	5.73	2.08	60.53	5.27
LK single	5.38	39.47	0.77	10.17	1.68	34.27	8.27
PCH double	3.12	9.07	16.88	20.08	3.38	36.4	11.07
PCH single	2.52	18.57	7.93	32.37	2.58	24.97	11.07
QLK	1.33	9.57	1.68	9.05	17.9	38.8	21.67
QR	1	10.17	0.9	8.57	1.23	52.37	25.77
QRE	1.67	7.3	1.3	5.37	2.13	16.7	65.53

TABLE XXXIV. 3×3 payoffs are scaled by 1/100 AIC winning probabilities.

DGP	# Parameters	LOOCV	2FCV	BIC	AIC
LK Double	3	22.63	38.5	47.57	47.13
LK Single	3	33.83	7.33	37.03	37.1
PCH Double	4	15.23	5.97	13.13	13.13
PCH Single	4	15.7	19.3	11.03	11.57
QLK	4	27.03	26.03	22.87	25.37
HQR	2	45.23	42.07	63.73	64
QRE	1	69.57	70.03	95.1	92.17

TABLE XXXV. Frequency with which each data generating process wins, excluding ties in Guessing games with fine strategy space

DGP	# Parameters	LOOCV	2FCV	BIC	AIC
LK Double	3	21.23	16.03	17.27	17.27
LK Single	3	19.37	13.27	16.27	16.27
PCH Double	4	21.97	13.2	5.1	5.1
PCH Single	4	20.87	11.3	5.13	5.13
QLK	4	4.83	4.03	0	0
HQR	2	3.5	4.07	0	0
QRE	1	0	0	0	0

TABLE XXXVI. Frequency with which each data generating process wins in a tie with at least one other model. in Guessing games with fine strategy space

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	QR	QRE
LK double	29.88	36.82	12.94	9.76	3.24	3.96	3.4
LK single	27.32	40.22	12.07	10.07	3.02	3.92	3.4
PCH double	22.81	16.58	23.01	22.16	3.41	4.56	7.47
PCH single	22.23	17.54	21.92	23.05	3.4	4.53	7.33
QLK	1.06	3.93	3.08	20.59	29.14	26.4	15.8
QR	1.81	14.99	1.62	7.44	10.21	46.61	17.33
QRE	12.53	11.02	0.45	2.13	1.26	3.03	69.57

TABLE XXXVII. Guessing Games fine strategy sets LOOCV winning probabilities.

DGP∣EST	LK double	LK single	PCH double	PCH single	QLK	QR	QRE
LK double	50.83	11.17	16.27	14.33	2.64	2.25	2.5
LK single	44.81	11.72	16.15	17.87	4.66	4.19	0.61
PCH double	34.23	5.04	12.5	29.89	4.2	2.26	11.87
PCH single	33.57	4.49	20.08	25.14	5.48	3.34	7.91
QLK	11.09	0	2.43	16.27	27.64	21.72	20.87
QR	8.28	0	1.64	10.73	13.07	43.87	22.4
QRE	20.99	1.1	0.12	1.93	2.65	3.18	70.03

TABLE XXXVIII. Guessing Games fine strategy sets 2FCV winning probabilities.

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	QR	QRE
LK double	56.2	36.6	1.93	2.17	0.73	0.83	1.53
LK single	47.2	45.17	2.18	1.82	0.9	1.13	1.6
PCH double	36.25	23.65	15.68	13.78	0.07	0.03	10.53
PCH single	34.53	25.07	15.7	13.6	0.07	0.07	10.97
QLK	0.3	3.5	0.13	1.27	22.87	46.57	25.37
QR	0.23	2.47	0.13	0.77	3.57	63.73	29.1
QRE	0.4	3.2	0	0.07	0	1.23	95.1

TABLE XXXIX. Guessing Games fine strategy sets BIC winning probabilities.

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	QR	QRE
LK double	55.77	36.5	2.3	2.73	0.8	0.77	1.13
LK single	46.87	45.23	2.52	2.32	0.97	0.87	1.23
PCH double	36.28	24.28	15.68	14.22	0.07	0.2	9.27
PCH single	34.6	25.53	15.7	14.13	0.1	0.23	9.7
QLK	0.37	4.4	0.2	2.17	25.37	45.53	21.97
QR	0.37	3.57	0.2	1.53	5.5	64	24.83
QRE	0.8	4.9	0	0.37	0.07	1.7	92.17

TABLE XL. Guessing Games fine strategy sets AIC winning probabilities.

DGP	# Parameters	LOOCV	2FCV	BIC	AIC
LK Double	3	18.37	35.83	72.8	72.67
LK Single	3	40.63	9.1	34.33	34.7
PCH Double	4	18.07	6.97	19.5	20.6
PCH Single	4	16.1	13.17	20.23	21
QLK	4	23.93	24.4	22.73	24.2
HQR	2	39.57	33.13	57.03	55.77
QRE	1	60.7	42.37	95.83	93.27

TABLE XLI. Frequency with which each data generating process wins, excluding ties in Guessing games with coarse strategy space

DGP	# Parameters	LOOCV	2FCV	BIC	AIC
LK Double	3	26.43	15.47	8.83	8.83
LK Single	3	19.4	22.37	8.3	8.3
PCH Double	4	22.63	18.07	6.77	6.77
PCH Single	4	18.83	20.93	6	6
QLK	4	7.4	6.8	0	0
HQR	2	5.3	4.67	0	0
QRE	1	0	0	0	0

TABLE XLII. Frequency with which each data generating process wins in a tie with at least one other model. in Guessing games with coarse strategy space

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	QR	QRE
LK double	29.35	39.51	10.5	8.19	3.6	5.81	3.03
LK single	23.94	48.24	8.97	8.58	3.18	3.69	3.4
PCH double	14.05	24.91	26.56	18.94	4.67	5.69	5.17
PCH single	12.54	26.16	23.45	23.18	4.16	5.32	5.2
QLK	9.59	13.33	6.09	8.23	26.6	23.65	12.5
QR	6.79	14.51	3.36	8.78	9.93	41.79	14.83
QRE	18.05	14.55	0.88	2.65	1.2	1.97	60.7

TABLE XLIII. Guessing Games coarse strategy sets LOOCV winning probabilities.

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	QR	QRE
LK double	57.02	16.95	12.07	7.89	2.02	1.52	2.53
LK single	43.84	19.17	23.16	7.7	2.65	2.7	0.78
PCH double	30.23	10.05	16.96	27.58	5.41	4.91	4.86
PCH single	23.57	9.62	32.74	23.12	5.81	4.38	0.76
QLK	9.83	8.56	1.57	5.71	27.64	25.71	20.98
QR	6.94	8.22	2.26	6.24	12.84	39.6	23.9
QRE	22.91	15.89	1	2.09	2.82	3.83	51.46

TABLE XLIV. Guessing Games coarse strategy sets 2FCV winning probabilities.

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	QR	QRE
LK double	77.22	16.22	0.77	0.83	1.07	1.4	2.5
LK single	53.98	38.48	0.8	1.03	0.9	1.83	2.97
PCH double	37.28	13.48	22.88	15.45	1.13	5.1	4.67
PCH single	31.88	18.92	15	23.23	0.93	4.77	5.27
QLK	2.33	5.93	0.83	1.53	22.73	41.8	24.83
QR	5.73	3.8	0.6	1.43	3.03	57.03	28.37
QRE	0.97	1.8	0.1	0.2	0	1.1	95.83

TABLE XLV. Guessing Games coarse strategy sets BIC winning probabilities.

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	QR	QRE
LK double	77.08	16.45	1.07	1.27	1.17	1.1	1.87
LK single	53.98	38.85	0.97	1.3	0.93	1.6	2.37
PCH double	36.62	13.72	23.98	16.32	1.23	4.77	3.37
PCH single	31.42	19.22	15.8	24	0.97	4.4	4.2
QLK	2.97	8.47	1.43	2.63	24.2	38.97	21.33
QR	6.4	6	1.1	2.2	4.33	55.77	24.2
QRE	1.43	3.1	0.47	0.27	0.07	1.4	93.27

TABLE XLVI. Guessing Games coarse strategy sets AIC winning probabilities.

Appendix F. Model Separability: Additional results

Model 1\Model 2	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
LK Double	77.5	88.7	82.7	89.9	89.9	70.5
LK Single		89.3	66.4	85.1	77.8	84.7
PCH Double			89.2	90.4	88.6	77.9
PCH Single				84.2	80.6	36.2
QLK					90.1	84.7
HQR						70.7

Table XLVII. Mean of $D_2(m1,m2;\theta)$ of 12 games' percentile in the distribution of randomly drawn 1000 games