Graz Economics Papers – GEP

GEP 2019–09

On the Predictive Power of Theories of One-Shot Play

Philipp Külpmann and Christoph Kuzmics

August 2019

Department of Economics
Department of Public Economics
University of Graz

On the Predictive Power of Theories of One-Shot Play*

Philipp Külpmann† Christoph Kuzmics‡

August 28, 2019

Abstract

We propose a novel challenge for assessing the predictive power of a theory of one shot-play in games (subjects playing a game exactly once): we test the predictive power of theories in situations for which we do not (yet) have any data. To do so, we consider a variety of such theories and fix their parameter estimates from the recent large scale meta-analysis of Wright and Leyton-Brown (2017). We then compare the predictive power of these theories, measured in terms of log-likelihood, for a series of symmetric hawk-dove games played in the lab.

We find that even for such a narrow class of games, no theory is uniformly better than all others across all treatments. Furthermore, the theory that provides the highest overall log-likelihood for our data is Nash equilibrium with risk aversion, with an estimated risk aversion parameter taken from Hey and Orme (1994) and its replication in Harrison and Rutström (2009). In particular, it significantly beats the two theories (based on quantal level k and cognitive hierarchy models) which performed best in Wright and Leyton-Brown's (2017) standard out-of-sample prediction task.

Keywords: hawk-dove games, testing theories, one-shot play, risk aversion, Nash equilibrium, quantal response equilibria, level-k theory, cognitive hierarchy theory

JEL codes: C72, C91

*Philipp is grateful for the hospitality of the economics department and the DR@W lab at the University of Warwick. Christoph is grateful to the economic theory group at CalTech for providing an inspiring setting to work on this paper in the winter and spring terms of 2018. We are grateful to Michael Greinecker, Christian Koch, Felix Mauersberger, Wieland Müller, Tom Palfrey, Karl Schlag, Daniel Sgroi as well as seminar audiences at Paderborn, Bielefeld, Innsbruck, and Vienna for helpful comments and suggestions, to James Wright for providing us the many parameter estimates, and to Hans Manner for pointing out and explaining the Vuong test to us. We acknowledge support through a grant from the German Research Foundation (DFG, grant number: KU 3071/1).

†Vienna Center for Experimental Economics, University of Vienna, philipp.kuelpmann@univie.ac.at

‡Department of Economics, University of Graz, christoph.kuzmics@uni-graz.at

Introduction Why do we try to develop a "theory" of human behavior? A good theory has a somewhat universal applicability (at least over some known domain of situations). It would, thus, allow analysts to make good predictions of behaviour for whatever situation (within the known domain) they are interested in without having to first resort to a time-consuming and costly experimental investigation. The purpose of this paper is to test whether existing theories for the specific setting of one-shot play make good predictions in this sense. With "one-shot play" we mean a situation in which subjects interact in a particular situation of strategic interaction (a game) exactly once in the lab.[1] Our research strategy here is to choose a class of situations of strategic interaction that is interesting in its own right, yet experimentally understudied, to assess the predictive power of various theories of one-shot play for this class. Importantly, for our research question, we have to estimate any parameter values for any theory we use with data from elsewhere. We cannot use any data from our own experiments for estimation.

In our analysis we, thus, strongly rely on the recent large-scale (in terms of games and data) and comprehensive (in terms of theories) meta-analysis of Wright and Leyton-Brown (2017). Wright and Leyton-Brown (2017) provide parameter estimates that best explain the data in terms of the (log-)likelihood for most well-known theories of one-shot play using data from an "extensive survey of papers", see (Wright and Leyton-Brown, 2017, Table 1), with a total of 13863 observations of behavior in a variety of games. The meta-analysis of Wright and Leyton-Brown (2017), by virtue of its comprehensiveness and our need for precisely estimated parameters, also essentially determines our choice of theories as exactly those theories that are analysed in Wright and Leyton-Brown (2017). We only add one additional theory: Nash equilibrium with risk aversion, where agents are assumed to have a CRRA utility function with a parameter of relative risk aversion, as recommended in Harrison and Rutström (2008), taken from Hey and Orme (1994) and its replication in Harrison and Rutström (2009).

We choose a set of symmetric hawk-dove games as our test-bed for these theories.[2] We do this for a variety of reasons. Hawk-dove games are simple in that there are only two players and two pure strategies. Taking into account symmetry restrictions (as imposed by our experimental design) they have a unique feasible Nash equilibrium and this Nash equilibrium is in completely mixed strategies.[3] Hawk-dove games are of independent interest as they are often used as the basic model of

[1] One should note that the analysts (we) cannot be sure that these subjects do not play this game or one that is "sufficiently" similar before or after the experiment as part of their interactions in their life.

[2] Note that there is no fully established terminology. What we here call hawk-dove games have also been called games of chicken or snowdrift. Some special cases have also been referred to as the battle of the sexes or anti-coordination games.

[3] This is so because we (completely) randomly match subjects and keep them completely anonymous from each other.

animal or human conflict, see e.g., the seminal work of Maynard Smith and Price (1973) and Maynard Smith (1982) as well as Baliga and Sjöström (2004) (with incomplete information), respectively. In special cases what we here call hawk-dove games have also been called battle-of-the-sexes or anti-coordination games which constitute a basic model of labor specialization or task allocation. See e.g., Rapoport (1966) and seminal applications to industrial organization in e.g., Farrell (1987), Farrell and Saloner (1985), and Dixit and Shapiro (1986). Finally, it seems that one-shot hawk-dove games are, despite all this, experimentally relatively understudied.[4]

Following Wright and Leyton-Brown (2017), we evaluate and compare our theories in terms of their log-likelihood. Our main results are as follows. First, we find that no theory is uniformly better than all others across all hawk-dove games.[5] Second, the theory that provides the highest overall log-likelihood for our hawk-dove data is Nash equilibrium with risk aversion. Third, and finally we find that only a quantal response equilibrium and slightly less so a noisy introspection model and a cognitive hierarchy model do not perform significantly worse than Nash equilibrium with risk aversion. In particular the two winning theories in Wright and Leyton-Brown (2017), a quantal level-k model proposed by Stahl and Wilson (1994) and a quantal cognitive hierarchy model proposed by Camerer, Nunnari, and Palfrey (2016), each with parameters taken from Wright and Leyton-Brown (2017), are significantly worse than Nash equilibrium with risk aversion.

In what follows we carefully describe the experimental design, list and the set of theories that we consider and state the main results, before we finally offer a discussion of our findings. Any details are given in the appendices.

Games and Experimental Design We have chosen a class of hawk-dove games for our main treatments. The payoff matrix for the hawk-dove game is given in Figure 1.

In all treatments we have that parameters $x, y \in \mathbb{N}$ with $x \geq 1$, $x > y$ and $y \geq 0$. Thus, all

Subjects are, thus, technically unable to coordinate on an asymmetric strategy profile. Appendix A provides a verbal explanation why this is so. See e.g., Kuzmics and Rogers (2010) or more generally Alos-Ferrer and Kuzmics (2013) for details as to how, why, and when symmetries restrict behavior in games.

[4]We are only aware of experiments, motivated by evolutionary game theory concerns, that study hawk-dove games (with and without anonymous matching) when subjects are repeatedly and randomly matched with each other to play the game over and over. See e.g., Oprea et al. (2011) and Benndorf et al. (2016). See Kuzmics and Rodenburger (forthcoming) for a an empirical analysis of a recurrent multiplayer game that also has hawk-dove features. There are also studies of repeated hawk-dove (or more specifically battle-of-the-sexes games), where the game is repeated with always the same set of players. See e.g., Kuzmics et al. (2014) and references therein.

[5]We also establish two unsurprising benchmark results. One, all theories have to be rejected at a high level of significance: no theory explains the data perfectly. Two, all theories except Nash equilibrium without risk-aversion provide significantly better predictions than uniformly random guessing.

	U	D
U	0, 0	x, 1
D	1, x	y, y

Figure 1: Payoff matrix: Hawk-dove game

hawk-dove games have the property that the unique best response to one pure strategy is the other pure strategy. The parameter choices for all treatments are given in Appendix B.

All such hawk-dove games have a unique symmetric Nash equilibrium, which is in properly mixed strategies. Also all such games of the matching pennies variety have a unique (typically asymmetric) Nash equilibrium, which again is in properly mixed strategies. The fact that hawk-dove games (and matching pennies games) produce properly mixed strategy predictions makes the statistical comparison between theories very clean and simple. This allows us to compare theories in terms of the (log-)likelihood.[6] One could also argue, by force of a Harsanyi (1973) purification argument, that mixed predictions are more robust to a small group of subjects doing strange things. The non-strange subjects can, in principle, counterbalance the strange behavior of a few by mixing appropriately so that aggregate play may be mixed in the way the theory predicts.

We also look at (asymmetric) games of the matching pennies variety, mostly to demonstrate that our subjects pool is not exceptional with details given in Appendix B. Matching pennies games have been widely studied experimentally. See e.g., Ochs (1995), Goeree and Holt (2001), and Goeree et al. (2003). The latter also uses risk aversion to improve the explanatory power of theories in these games.

We asked 147 subjects, in a random matching environment, to choose actions once each for 10 different hawk-dove games and once each for 10 different matching pennies games. Subjects get no feedback about their opponents' action choices and the order in which they make their choices for the various games is random. The details of the experimental design can be found in Appendix B.

Theories We then need to identify the class of theories that we want to compare. One problem here is that many theories have parameters that, for our purpose, need to be fixed before we take them to our data. Here we have the good fortune of having access to a recent large-scale and comprehensive meta-analysis of exactly these theories in Wright and Leyton-Brown (2017) which is where we take all parameter estimates for these theories from. These theories include **Nash equilibrium (NE)** theory, a **level-k reasoning (LK)** model as in Stahl and Wilson (1994), Stahl and Wilson (1995) and Nagel (1995), a **cognitive hierarchy (CH)** model of Camerer, Ho, and Chong (2004), a **quantal**

[6] Any mixed data under the null hypothesis of a deterministic prediction has an exactly zero likelihood, while any mixed prediction will provide a positive likelihood in every case.

response equilibrium (QRE) of McKelvey and Palfrey (1995), a **noisy introspection (NI)** model of Goeree and Holt (2004), a **quantal level-k (QLK)** model as proposed by Stahl and Wilson (1994), and a **quantal cognitive hierachy (QCH)** model of Camerer, Nunnari, and Palfrey (2016). The last two are the two "winning" theories from the Wright and Leyton-Brown (2017) meta-analysis. In additional to all these we consider **Nash equilibrium with risk aversion (NERA)** (which is not considered in Wright and Leyton-Brown (2017)), where we use an estimated coefficient of relative risk aversion (for a CRRA utility function) as recommended in Harrison and Rutström (2008) from Hey and Orme (1994) and its replication in Harrison and Rutström (2009).[7] Finally, we have added a **benchmark "theory" (RND)** in which every player picks each action with a probability of 50%. All the theories have in common that they are giving us different predictions in mixed strategies. For a detailed description of all theories considered here see Appendix C.

We have excluded theories which predict pure strategies in some treatments as they produce minimal likelihood values of zero and, hence, do not perform well given our evaluation criteria of likelihood-based predictive power. These include maximin play, but also "level 1 with risk aversion" as identified in Fudenberg and Liang (forthcoming) as in hawk-dove games this theory also makes pure strategy predictions.[8] We have excluded theories whose predictions are identical to those of other theories. These include minimax regret theory as its predictions are the same as Nash equilibrium predictions in hawk-dove games. The same is true for theories of Nash equilibrium with a fraction of fairness-minded individuals as put forward in Fehr and Schmidt (1999) as well as for theories with a fraction of individuals who care for relative payoffs as proposed by Bolton and Ockenfels (2000). In our hawk-dove games the predictions of these theories are again equal to Nash equilibrium predictions. This is so because of the Harsanyi (1973) purification argument that we described in the previous section: the fraction of non-fairness-minded individuals is large enough (in the respective calibrations) to "equilibrate" play. We have, finally, also ignored some of the theories developed (and calibrated) for stationary long-run behavior such as those discussed and experimentally compared in e.g., Selten and Chmura (2008), with comments by Brunner et al. (2010), and a reply by Selten et al.

[7] Risk aversion is also used to improve predictive power in the machine-learning approach of Fudenberg and Liang (forthcoming). There risk aversion is used not for Nash equilibrium, but for the behavior of the level 1 individuals in a level-k model. Risk-aversion has been used to explain behavior in experiments in a variety of studies. For instance Goeree and Holt (2004), Goeree, Palfrey, and Holt (2003), and Fudenberg and Liang (forthcoming) find that risk-aversion improves the explanatory power of equilibrium behavior in matrix games, while Cox and Oaxaca (1996), Chen and Plott (1998), Goeree, Holt, and Palfrey (2002), and Campo, Guerre, Perrigne, and Vuong (2011) find that risk-aversion is helpful in explaining behavior in auctions.

[8] We also do not look at the machine-learning optimized algorithm that Fudenberg and Liang (forthcoming) identify because it applies only to three strategy games. One could understand our two strategy games as three strategy games with a dominated strategy, but we did not pursue this.

(2011).

Results A theory i makes prediction $p_{i,t}$ for treatment t, where $p_{i,t}$ is the proportion of hawk (in hawk-dove games) or the proportion of heads (in matching pennies style games). For our purposes, theory i is thus identified by the vector $p_i = (p_{i,1}, ..., p_{i,20})$ of predictions. These predictions of the various theories as well as the observed proportions of hawk or heads are provided in Table 2 for hawk-dove games and Table 3 for matching pennies games in Appendix D. Let p denote the true probability vector of choices. We first establish two benchmark results.

Result 0.

a) For each considered theory i we reject the null of this theory making correct predictions ($p = p_i$) at a high level of significance.

b) All theories except Nash equilibrium without risk-aversion (NE) are significantly better than RND.

In order to compare theories we compute their (overall) log-likelihoods (given the data). The full table of log-likelihood values for all theories and all treatments as well as the overall log-likelihood values for all theories are given in Table 4 in Appendix D. One finding stands out:

Result 1. *The theory with the highest overall log-likelihood for hawk-dove games among all considered theories is Nash equilibrium with risk aversion.*

Figure 2 plots the observed and predicted frequencies of the action "hawk" for the four best theories (that can be shown to be all overall significantly better than all other considered theories - see below) across all hawk-dove treatments.

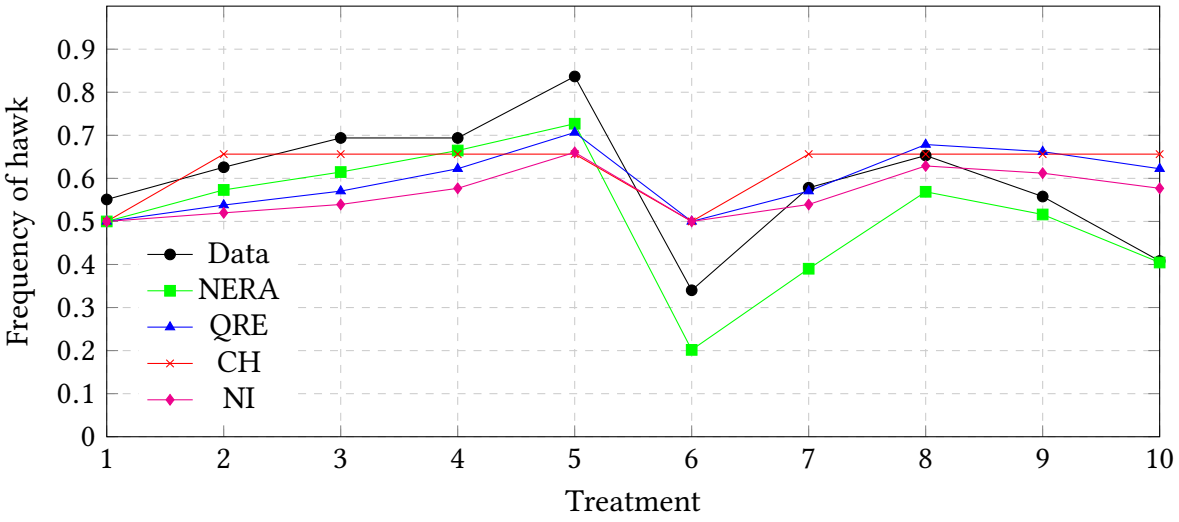

Figure 2: Observed and predicted frequencies of action hawk in hawk-dove games

This figure suggests and this is substantiated by the appropriate test (see below) the following:

Result 2. *Among all considered theories none provides universally best predictions in all hawk-dove games.*

For instance, the Voung z-score (see more below) for comparing the two top theories NERA and QRE for treatment 10 is 2.04 and for treatment 7 is −2.88.[9] This means that even the small class of hawk-dove games is sufficiently varied so that no theory is universally better within this class than all other theories. This can also be seen somewhat in Figure 2 which plots the predictions of the four best theories in our hawk-dove games in comparison to the actual data. It is probably best seen in Figure 3 which plots the log-likelihood of the best four theories in our hawk-dove games across treatments.

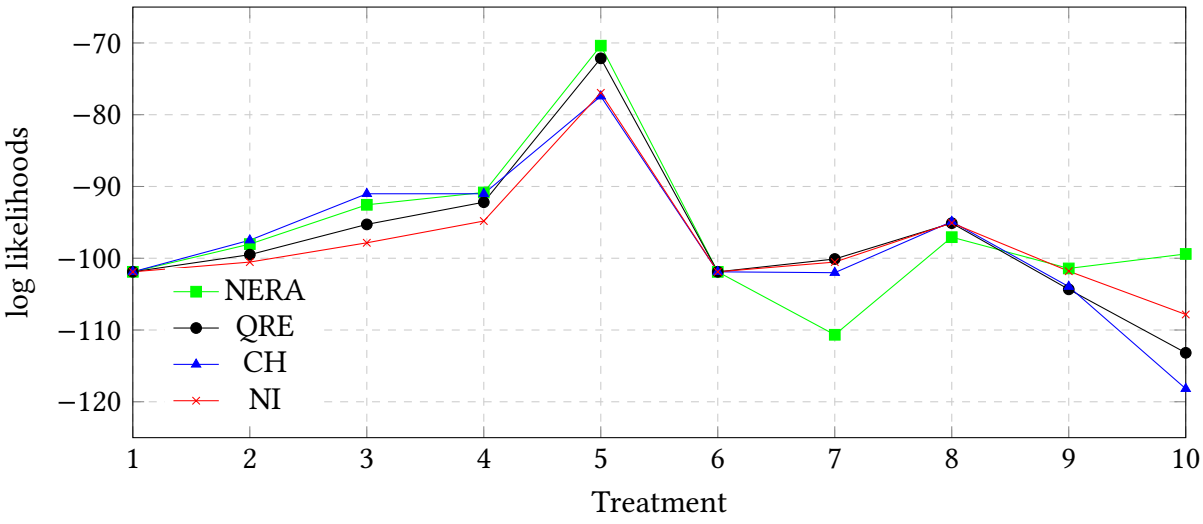

Figure 3: Log likelihoods of the contenders in the hawk-dove games

To then compare any two theories statistically, essentially knowing that both are wrong, we can perform a Vuong (1989) test of the null that both theories are equally far from the "true" theory in terms of the expected log-likelihood relative to the true theory – this is the Kullback-Leibler divergence (or relative entropy). The log-likelihood of all theories for all treatments are given in Table 4 in Appendix D, and the Vuong test is described in Appendix D.3. The main findings can be summarized as follows:

Result 3. *Some theories predict the behavior in the hawk-dove games significantly better than others. The contenders for the best theory in hawk-dove games are Nash equilibria with risk aversion (NERA), quantal response equilibria (QRE), and a little worse (but not statistically significantly so) cognitive hierarchy (CH) and noisy introspection (NI).*

[9] I.e., QRE is significantly closer to the true theory than NERA in treatment 7 and significantly further away from the true theory in treatment 10. See below or in Appendix D.3 for more.

Table 1 provides the Vuong z-scores for hawk-dove games of any (row) theory when compared to any (column) theory. A negative z-score means that the likelihood of the row theory is below that of the column theory (i.e., it is better), while a positive z-score means that the likelihood of the row theory is above that of the column theory (i.e., it is worse).[10]

	NE	NERA	LK	CH	QRE	QLK	QCH	RND	NI
NE	0	5.86	4.31	7.98	7.73	9.3	5.76	3.49	6.47
NERA	-5.86	0	-3.71	-1.21	-1	-1.57	-2.05	-4.98	-1.43
LK	-4.31	3.71	0	3.05	4.35	1.76	6.89	-6.49	6.13
CH	-7.98	1.21	-3.05	0	0.92	-1.04	-0.82	-3.77	0.1
QRE	-7.73	1	-4.35	-0.92	0	-2.42	-2.34	-5.05	-1.27
QLK	-9.3	1.57	-1.76	1.04	2.42	0	0.07	-2.57	0.97
QCH	-5.76	2.05	-6.89	0.82	2.34	-0.07	0	-7.62	4.32
RND	-3.49	4.98	6.49	3.77	5.05	2.57	7.62	0	6.81
NI	-6.47	1.43	-6.13	-0.1	1.27	-0.97	-4.32	-6.81	0

Table 1: Vuong z-scores: HDG

Among the results mentioned above, Table 1 also shows that our benchmark "theory" of uniformly random behavior, was outperformed significantly by every theory except Nash equilibrium. Thus, while no theory can explain the observed results, almost every theory explains behavior better than just guessing. Furthermore, the two theories which performed best in Wright and Leyton-Brown (2017), i.e., the quantal level-k and quantal cognitive hierarchy theories, are worse than the best two theories. This is also true in the matching pennies games in which neither of these two theories is among the significantly best theory, see Table 6 in the Appendix.

Discussion The main result of this paper, in our mind, is Result 2. It demonstrates that even for such a small set of only hawk-dove games none of the theories of one-shot play considered here (and calibrated with data from other games) is universally better than all other theories in terms of their predictive power. Thus, while for each theory there may be games for which this theory provides good predictions of one-shot play, we can also always find games for which it does not. Moreover it seems to us that we still do not have a good understanding as to which features of a game make which theory relevant (or lead to which suitable parameter choice for a given theory).

[10] In the table, 5% level significant positive z-scores (i.e., a score above 2 which means that the row theory is significantly worse) are marked in red, and 5% level significant negative scores (i.e., below −2) are marked in green. "Suggestive" z-scores between 1 and 2 (−1 and −2) are colored in light red (light green).

Somewhat interestingly we found that, for the class of games that we chose for our testing exercise, the simple theory of Nash equilibrium, albeit under the assumption of risk averse subjects, explains the data altogether at least as well as any other theory, and better than most. Note that there are no good a priori reasons, neither epistemic nor evolutionary, that suggest Nash equilibrium to be a good theory of one-shot play for hawk-dove games. Even under the here implausible epistemic assumption of a common knowledge of rationality among the players, we at best only expect any rationalizable outcome (see e.g., Bernheim (1984) and Pearce (1984)), and for hawk-dove games this is the set of all strategy profiles. Evolutionary arguments (see e.g., the "mass atom" interpretation of Nash equilibrium in Nash (1950) and Weibull (1995) and Sandholm (2010) for a textbook treatment) do not apply because each game is played only once by every subject, no learning or evolutionary adjustment can occur (unless subjects have played this game often before, and recognize this, which seems unlikely).[11] One could possibly argue that the symmetry in hawk-dove games makes it relatively easy for subjects to put themselves in their opponent's shoes, as their opponent's payoffs and options are exactly the same as their own, and that this could possibly make Nash equilibrium (in mixed strategies) more likely. While we, thus, find this result - that Nash equilibrium under risk aversion explains the data best - somewhat surprising and interesting, we do not believe that it would necessarily hold had we chosen another set of previously understudied games for our exercise. For instance, note that at least for games of the matching pennies variety (which are not understudied) Nash equilibrium even with risk aversion does not overall produce the best predictions. One should also note, however, that in all our matching pennies treatments one player position's behavior is actually also best explained by Nash equilibrium with risk aversion: the player position that obtains symmetric payoffs from both actions and, in Nash equilibrium has to balance her two actions in a non-trivial way to make her opponent indifferent. See Figure 5 in the Appendix. Risk aversion was not considered in Wright and Leyton-Brown (2017) and we conjecture that the QRE parameter would have changed had risk aversion been estimated at the same time. Nevertheless we conjecture that if

[11]There is a fair amount of evidence that playing a game sufficiently often (recurrently with different opponents) does lead to equilibrium play, while playing a game only once does not deliver Nash equilibrium play. For instance, Van Huyck, Battalio, and Beil (1990) show that subjects often fail to play any Nash equilibrium in one-shot coordination games, while Cooper, DeJong, Forsythe, and Ross (1990) find evidence of Nash equilibrium play in recurrent coordination games. Similarly O'Neill (1987) finds evidence against laboratory subjects playing minmax (i.e. Nash equilibrium) strategies in zero-sum games, while Walker and Wooders (2001) find mixed evidence that professional tennis players use minimax strategies in their service game, Hsu, Huang, and Tang (2007) find evidence that professional tennis players use minimax strategies, and Palacios-Huerta (2003) finds strong evidence that professional soccer players (and goalkeepers) use minimax strategies when taking (or defending) penalty kicks. Professionals have played these games often, while laboratory subjects not (or not often enough). Binmore, Swierzbinski, and Proulx (2001) find that after and only after sufficient practice with the game do lab subjects play minimax strategies in a set of zero-sum games.

we had had an appropriately estimated QRE model with risk aversion, it would possibly outperform the Nash equilibrium with risk aversion model. See e.g., Goeree, Holt, and Palfrey (2002) and Bhattacharya, Duffy, and Kim (2017) for QRE with risk aversion. The level-k and quantal level-k models would not produce very different predictions for hawk-dove games even if we consider risk aversion.

There is also a simple argument to be made that no simple theory as we considered them in this paper can even hope to explain one-shot behavior in all games. For instance, Burnham et al. (2009) found that in beauty contest games a player's action choice correlates with her cognitive ability. If this is true, presumable a group of high cognitive ability players playing this game would make statistically distinguishable aggregate choices than a group of lower cognitive ability players. All the theories that we consider, at least as we consider them in this paper, as they do not depend on the cognitive ability characteristics of the players, do not distinguish between these two situations and would therefore make the same prediction in both cases, which would have to be false in at least one of the two situations.

APPENDIX

A On symmetric play in symmetric hawk-dove games

Suppose we have one large population of individuals with varying characteristics (in terms of sex, age, et cetera, as it is in our subject pool) who are randomly matched to play a hawk-dove game. Compare two situations. One, as in our experimental design, the two players learn nothing about each other's characteristics (as nobody knows whom they are playing against). Two, and not in our design, the two players observe each other before they play. In the latter case one can readily imagine that subjects use the information they get to help them play an asymmetric (equilibrium) outcome. For instance, if it emerges that one player is a woman and the other a man, we can have that the woman plays U and the man D, which given the observability of each other's characteristics would lead to an asymmetric outcome. Now if it emerges that both players are of the same sex, then they cannot condition on that information to achieve an asymmetric outcome. They could then perhaps condition on the two player's relative age. As long as characteristics are sufficiently varied and publicly observed players could thus play an asymmetric outcome. But note that it is crucial here that players not only condition on their own characteristics but also on their opponents to achieve this asymmetric outcome. A young woman may have to play U against any man, but D against an older woman. If we remove the information about the opponent's characteristics this grand strategy that guarantees asymmetric outcomes is no longer feasible. Every player can still condition on her

own characteristics, for instance every young woman can play U while every old man plays D, but as they are not necessarily matched with each other the overall outcome cannot be asymmetric in all matches. In the end there will be a proportion of individuals playing U (perhaps because, as in a Harsanyi (1973) purification story, they condition on their own private type) and a remaining proportion of individuals playing D and because of the random matching the overall proportion of asymmetric outcomes simply must be the product of the two proportions (of U and D). We can, thus, restrict attention to symmetric play (with perhaps a purification argument in the back of our mind).

B Experimental Design

The experiment was conducted at the DR@W Laboratory at the University of Warwick using zTree (Fischbacher (2007)). A total of 147 subjects played 10 different hawk-dove games and matching pennies games each. No game was played more than once. Whether hawk-dove or matching pennies was played first was determined randomly and so was the order of the 10 types for each game within each class.

	U	D
U	0, 0	x, 1
D	1, x	y, y

(a) HDG

	U	D
U	x, 0	0, 1
D	0, 1	1, 0

(b) MP

Figure 4: Payoff tables

For 5 rounds in the *hawk-dove games*, subjects played a pure anti-coordination game (battle-of-the-exes), i.e., $y = 0$ (Figure 4(a)) and $x \in \{1, 2, 3, 5, 10\}$. The other 5 rounds are variations of the classical hawk-dove game with $(x, y) \in \{(3, 2)(5, 2)(10, 2)(10, 3)(10, 5)\}$. At the very end of the experiment, one of the 10 rounds was randomly selected and paid out in *GBP*.

For the *matching pennies games* the 10 rounds consisted of playing the game depicted in Figure 4(b) with $x \in \{1, 2, 3, 5, 10\}$ each once as player 1 and 2 in a random order. Again, at the end of the experiment, one of the 10 rounds was randomly selected and paid out in *GBP*.

Subjects were for each round randomly matched with some other subject in the subject pool. Subjects never received any feedback about their opponent or their opponent's strategy choice until the very end when all they were told is how much money they received.

After the games were played, we elicited risk aversion and level-k reasoning skills (in the 11/20 game developed by Arad and Rubinstein (2012)) which were not used in this paper.

Before and during the experiment we had a number of instructions (on paper, read out loud and on the screen) and quizzes.

For the instructions, the zTree code of the experiment, the R code which was used to generate the model predictions and to run the tests, the data and any additional information please contact the authors.

C Theories

C.1 Risk Aversion

We are using a CRRA utility function of the following form:

$$u_{CRRA}(x) = \frac{x^{1-\rho}}{1-\rho}$$

The parameter ρ we have taken from Hey and Orme (1994) and its replication as reported and recommended in Harrison and Rutström (2008).

They used an extensive random lottery pair design in which they asked subjects to make choices between lotteries using 4 fixed prices and varying probabilities. Fortunately, their results are robust in the payment domain we are using them in and also across different countries and currencies, as shown by Harrison and Rutström (2009) and Harrison and Rutström (2008, p121-122).

C.2 Level-k reasoning

The prediction of level-k reasoning models depends on two parameters: level-0 behavior and the distribution of levels among the players. Usually, mixing $50-50$ is assumed to be the natural level-0 assumption. However, the experimental design supports U to be the natural level-0 choice.

If we assume that a level-0 player plays U, every even level player plays U and every odd level player plays D. If we assume that a level-0 player plays $50-50$ (or D) or is mixing, every even level player plays D and every odd level player plays U.[12] Thus, the predictions for the HDG only depend on the distribution of levels (let's call $prop_{even}$ the proportion of even level players) and is, assuming the level-0 to be U,

$$p_1 = p_2 = prop_{even}$$

[12]The only exception of this is when the level-0 player is assumed to play $50-50$ and $50-50$ is a Nash equilibrium (i.e., a fixed point of the best responses) for both players. Then the best response of every player to $50-50$ will always be $50-50$ and so one. This is the case for treatments 1 and 6 in both classes of games.

and for level-0 D or mixing just $prop_{odd}$. For matching pennies it depends on the assumption of level-0 of both players. We get for level-0 being U:

$$p_1 = prop_{even}, p_2 = prop_{odd}$$

or vice versa for the other level-0 beliefs.

We have taken the type distribution from Arad and Rubinstein (2012, p. 3566, footnote 6).

Remark (Structure of predictions)**.** *Note that strategy choices are independent of x and y and only depend on the level of the player (except in the case mentioned in Footnote 12).*

C.3 (Poisson) Cognitive Hierarchy

Definition 1 (Poisson) Cognitive Hierarchy (from Wright and Leyton-Brown (2017)))**.** Let $\pi_{i,m} \in \Pi(A_i)$ be the distribution of actions predicted by agent i with level m by the Poisson-CH model. Let $f(m) = \text{Poisson}(m;\tau)$. Let $BR_i^G(s_{-i}$ denote the set of i's best responses in game G to the strategy profile s_{-i}. Let

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

be the truncated distribution of actions predicted for an agent conditional on that agent's having level $0 \leq l \leq m$. Then π is defined as

$$\pi_{i,0} = |A_i|^{-1}$$

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

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

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

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

This parameter we have taken, again, from Wright and Leyton-Brown (2017).

Remark (Structure of predictions)**.** *Due to the special structure of the level-k reasoning predictions, the predictions of cognitive hierarchy are also independent of the payoffs. Again, as in the case of level-k reasoning, Footnote 12 also applies here.*

C.4 Noisy Introspection

We are using the version of noisy introspection as proposed by Goeree and Holt (2004) and as defined in Wright and Leyton-Brown (2017, Definition 6):

Definition 2 (NI model (Wright and Leyton-Brown))**.** Define $\pi_{i,k}^{NI,n}$ as

$$\pi_{i,k}^{NI,n} = \begin{cases} QBR_i^G\left(\pi_{-i,k+1}^{NI,n}; \frac{\lambda_0}{t^k}\right) & \text{if } k < n, \\ QBR_i^G\left(\pi_0; \frac{\lambda_0}{t^k}\right) & \text{otherwise,} \end{cases}$$

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

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

C.5 Quantal Responses and Quantal Responses Equilibria

A logit quantal response $QBR_i(s_{-i}, \lambda)$ of player i is a reaction to the strategy profile s_{-i}, s.t.:

$$s_i(a_i) = \frac{exp(\lambda u_i(a_i, s_{-i}))}{\sum_{\forall a' \in A} exp(\lambda u_i(a_{i'}, s_{-i}))}$$

Quantal response equilibrium is, like Nash, an equilibrium concept, i.e., it assumes that every player's strategy is a best response to the strategy of the other player, i.e., $p_i^* = QBR(p_j^*, \lambda)$ and $p_j^* = QBR(p_i^*, \lambda)$.

Quantal responses is not invariant to scaling, i.e., the results depend on the scaling of payments as already pointed out by Wright and Leyton-Brown (2010).

We have chosen the same scaling as Wright and Leyton-Brown (2017) did, i.e., we normalized the payments to expected (USD) cents.[13]

C.6 Quantal Level-k

The second to last theory we are considering is a model of Quantal level-k as suggested by Wright and Leyton-Brown (2017):

We are restricting the model to 4 levels (i.e., the max level is 3) with homogeneous precision but general beliefs about the precision of others.

[13] As the experiment was run in the UK, we had to fix the exchange rate from GBP to USD and we decided to fix it at 1.41, which is the rounded, weighted (by subjects or sessions) average of exchange rates on the days the experiment was run.

Therefore, we have 7 parameters, 4 precision parameters: the real precision parameter for all types: λ, the perceived prevision parameter level-2 has about level-1, the perceived precision parameter level-3 has about level-2 and lower and the perceived precision parameter level-3 thinks level-2 has about level-1.

Furthermore, we have 3 parameters for the proportion of level-1,2 and 3 players (and the rest being level-0).

Let's call the probability distribution of player i with level j, $p_{i,j}$ over actions a_i

$$\begin{aligned} p_{i,0}(a_i) &= |A_i|^{-1} = \frac{1}{2} \\ p_{i,1} &= QBR_i(p_{-i,0}, \lambda) \\ p_{i,1(2)} &= QBR_i(p_{-i,1}, \lambda_{1(2)}) \\ p_{i,2} &= QBR_i(p_{-i,1(2)}, \lambda) \\ p_{i,1(2(3))} &= QBR_i(p_{-i,1}, \lambda_{1(2(3))} \\ p_{i,2(3)} &= QBR_i(p_{-i,1(2(3))}, \lambda_{2(3)}) \\ p_{i,3} &= QBR_i(p_{-i,2(3)}, \lambda) \end{aligned}$$

where $p_{i,1(2)}$ is the mixed strategy profile representing level-2 player's prediction of how players 1 and 2 will play, $p_{i,2(3)}$, level-3's prediction of level 2 players and $p_{i,1(2(3))}$ level-3's prediction of level-2's prediction of level-1 players.

Again, parameters were taken from Wright and Leyton-Brown (2017).

C.7 Quantal Cognitive Hierarchy

The last theory was also suggested by Wright and Leyton-Brown (2017): Logit quantal cognitive hierarchy with homogeneous and accurate beliefs.

This is a version of cognitive hierarchy (Appendix C.3) but instead of best responses $BR_i(\cdot)$ logit quantal best responses, $QBR_i(\cdot; \lambda)$ as in Appendix C.5 are used.

Thus, this theory has two parameters λ and τ which were taken from Wright and Leyton-Brown (2017).

D Results

D.1 Predictions, Data, and Log-likelihood

Table 2 provides the observed frequency of hawk, the parameters (x and y) describing each treatment, as well as all predictions of this frequency of the various theories for all hawk-dove treatments. Table 3

provides the analogue of this for the matching pennies treatments.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Data	0.55	0.63	0.69	0.69	0.84	0.34	0.58	0.65	0.56	0.41
x	1.00	2.00	3.00	5.00	10.00	3.00	5.00	10.00	10.00	10.00
y	0.00	0.00	0.00	0.00	0.00	2.00	2.00	2.00	3.00	5.00
NE	0.50	0.67	0.75	0.83	0.91	0.50	0.75	0.89	0.88	0.83
NERA	0.50	0.57	0.61	0.66	0.73	0.20	0.39	0.57	0.52	0.40
LK	0.50	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53
CH	0.50	0.66	0.66	0.66	0.66	0.50	0.66	0.66	0.66	0.66
QRE	0.50	0.54	0.57	0.62	0.71	0.50	0.57	0.68	0.66	0.62
QLK	0.50	0.55	0.59	0.67	0.79	0.50	0.59	0.76	0.74	0.67
QCH	0.50	0.52	0.53	0.56	0.62	0.50	0.53	0.60	0.59	0.56
RND	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
NI	0.50	0.52	0.54	0.58	0.66	0.50	0.54	0.63	0.61	0.58

Table 2: Predictions: HDG

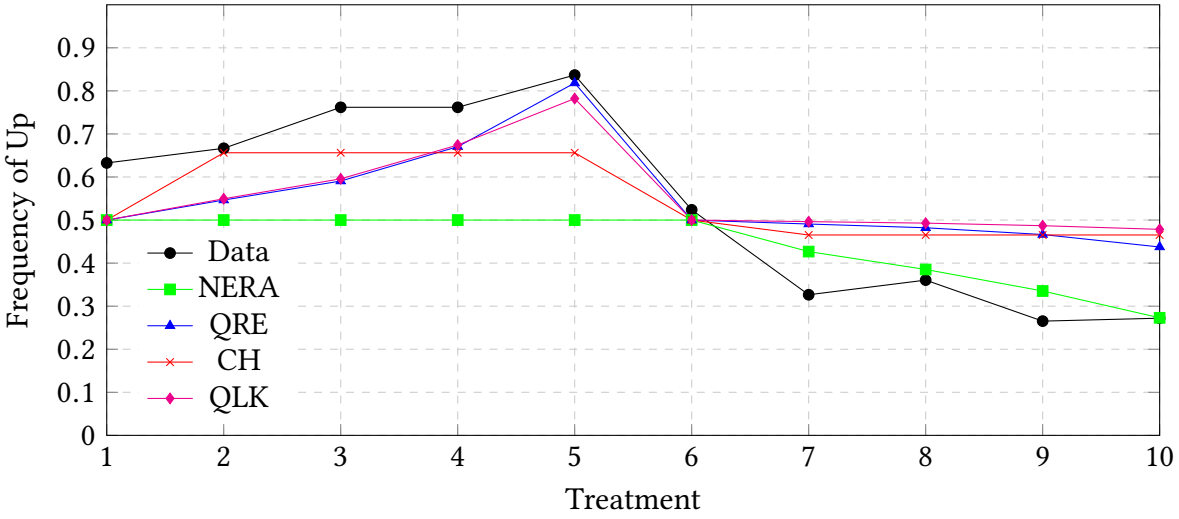

Figure 5: Observed and predicted frequencies of action up in the matching pennies games

Table 4 provides the log-likelihood of all theories for all treatments and in sum for the hawk-dove games as well as the matching pennies games.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Data	0.63	0.67	0.76	0.76	0.84	0.52	0.33	0.36	0.27	0.27
x	1.00	2.00	3.00	5.00	10.00	1.00	2.00	3.00	5.00	10.00
y	1.00	1.00	1.00	1.00	1.00	2.00	2.00	2.00	2.00	2.00
NE	0.50	0.50	0.50	0.50	0.50	0.50	0.33	0.25	0.17	0.09
NERA	0.50	0.50	0.50	0.50	0.50	0.50	0.43	0.39	0.34	0.27
LK	0.50	0.53	0.53	0.53	0.53	0.50	0.47	0.47	0.47	0.47
CH	0.50	0.66	0.66	0.66	0.66	0.50	0.47	0.47	0.47	0.47
QRE	0.50	0.55	0.59	0.67	0.82	0.50	0.49	0.48	0.47	0.44
QLK	0.50	0.55	0.60	0.67	0.78	0.50	0.50	0.49	0.49	0.48
QCH	0.50	0.52	0.53	0.56	0.63	0.50	0.50	0.50	0.50	0.50
RND	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
NI	0.50	0.52	0.54	0.59	0.69	0.50	0.55	0.66	0.64	0.60

Table 3: Predictions: Matching Pennies

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	SUM
NE	-101.89	-97.73	-91.73	-99.23	-69.27	-101.89	-110.40	-123.37	-146.11	-166.82	-1108.44
NERA	-101.89	-98.03	-92.55	-90.83	-70.38	-101.90	-110.66	-97.06	-101.42	-99.40	-964.13
LK	-101.89	-99.94	-98.73	-98.73	-96.21	-104.98	-100.78	-99.45	-101.14	-103.78	-1005.63
CH	-101.89	-97.48	-91.01	-91.01	-77.43	-101.89	-102.01	-94.90	-103.95	-118.19	-979.78
QRE	-101.89	-99.50	-95.28	-92.19	-72.14	-101.89	-100.10	-95.11	-104.33	-113.18	-975.63
QLK	-101.89	-99.02	-93.73	-90.67	-66.42	-101.89	-100.16	-99.10	-111.82	-121.28	-985.98
QCH	-101.89	-100.80	-98.64	-96.11	-82.15	-101.89	-100.75	-95.85	-101.14	-106.19	-985.42
RND	-101.89	-101.89	-101.89	-101.89	-101.89	-101.89	-101.89	-101.89	-101.89	-101.89	-1018.93
NI	-101.89	-100.53	-97.85	-94.82	-76.96	-101.89	-100.53	-95.08	-101.81	-107.84	-979.20

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	SUM
NE	-101.89	-101.89	-101.89	-101.89	-101.89	-101.89	-92.87	-100.52	-89.57	-106.11	-1000.43
NERA	-101.89	-101.89	-101.89	-101.89	-101.89	-101.89	-95.97	-96.29	-86.72	-86.05	-976.38
LK	-101.89	-99.21	-97.53	-97.53	-96.21	-101.89	-99.09	-99.69	-98.01	-98.13	-989.21
CH	-101.89	-93.60	-84.55	-84.55	-77.43	-101.89	-98.71	-99.40	-97.46	-97.60	-937.08
QRE	-101.89	-97.96	-90.24	-83.63	-65.59	-101.89	-100.98	-100.52	-97.58	-94.63	-934.92
QLK	-101.89	-97.75	-89.70	-83.40	-66.79	-101.89	-101.53	-101.32	-100.14	-99.14	-943.57
QCH	-101.89	-100.39	-97.25	-93.33	-80.91	-101.89	-101.86	-101.85	-101.74	-101.58	-982.70
RND	-101.89	-101.89	-101.89	-101.89	-101.89	-101.89	-101.89	-101.89	-101.89	-101.89	-1018.93
NI	-101.89	-99.96	-95.88	-90.76	-73.46	-101.87	-107.14	-122.80	-127.14	-117.78	-1038.68

Table 4: Loglikelihoods: HDG & MP

D.2 Testing Theories Individually

Let $\bar{p}_t$ be the empirical proportion of hawk in treatment t and let $p_{i,t}$ be the theoretical proportion in treatment t according to theory i. Let $z_{i,t} = \left(\bar{p}_t - p_{i,t}\right) / \sqrt{\left(p_{i,t}(1-p_{i,t})/n\right)}$. Then, by the usual central limit theorem argument, $z_{i,t}$ is asymptotically standard normally distributed under the null that theory i is correct, i.e., that the true $p_t = p_{i,t}$. Let $\chi_i^2 = \sum_{t=1}^{10} z_{i,t}^2$. Then χ_i^2 is asymptotically chi-squared distributed with 10 degrees of freedom. Table 5 reports the p-values of this test of $p = p_i$ (recall the vector notation) for each theory i.

	chi-square(HDG)	p-value(HDG)	chi-square(MP)	p-value(MP)
NE	482.47	<0.00001	252.66	<0.00001
NERA	61.2	<0.00001	184.01	<0.00001
LK	134.97	<0.00001	207.75	<0.00001
CH	90.65	<0.00001	110.09	<0.00001
QRE	81.78	<0.00001	107.79	<0.00001
QLK	109.56	<0.00001	124.63	<0.00001
QCH	97.02	<0.00001	197.2	<0.00001
RND	161.05	<0.00001	266.41	<0.00001
NI	86.14	<0.00001	322.75	<0.00001

Table 5: Testing individual theories

D.3 Vuong Test

The idea utilized by Vuong (1989) is that there is a "true" theory $p = (p_1, p_2, ..., p_{10})$. Let $\bar{p}_t$ denote the observed proportion of hawk or heads in treatment t (of the hawk-dove or matching pennies games) among the $n = 147$ subjects. Let $\bar{p} = (\bar{p}_1, ..., \bar{p}_{10})$. Then the log-likelihood ratio between any two theories i and j is given by

$$\log LR(\bar{p}, p_i, p_j) = \sum_{t=1}^{10} \left[\bar{p}_t \log\left(p_{i,t}/p_{j,t}\right) + (n - \bar{p}_t) \log\left((1-p_{i,t})/(1-p_{j,t})\right)\right].$$

The "true" variance of this log-likelihood is then given by

$$\sum_{t=1}^{10} np_t(1-p_t)\left[\log\left(p_{i,t}/p_{j,t}\right) - \log\left((1-p_{i,t})/(1-p_{j,t})\right)\right]^2,$$

which can be estimated by replacing p_t by its maximum likelihood estimator $\bar{p}_t$ for each treatment t.

A Vuong statistic (or z-score) can then be computed as the log-likelihood divided by the square root of its estimated variance. This statistic, under the true model, is asymptotically standard normally distributed by the usual central limit theorem argument.[14]

Whether or not a theory describes the behavior in our experiment better or worse than another theory can thus be measured in terms of the Vuong z-score. Table 1 provides the Vuong z-scores for hawk-dove games of any (row) theory when compared to any (column) theory. A negative z-score means that the likelihood of the row theory is below that of the column theory (i.e., it is better), while a positive z-score means that the likelihood of the row theory is above that of the column theory (i.e., it is worse). In the table, 5% level significant positive z-scores (i.e., a score above 2 which means that the row theory is significantly worse) are marked in red, and 5% level significant negative scores (i.e., below −2) are marked in green. "Suggestive" z-scores between 1 and 2 (−1 and −2) are colored in light red (light green). The results for the matching pennies games are given in Table 6.

	NE	NERA	LK	CH	QRE	QLK	QCH	RND	NI
NE	0	2.48	0.71	3.76	3.79	3.19	1.04	-1.1	-1.72
NERA	-2.48	0	-2.1	4.45	4.25	3.35	-0.83	-5.93	-4.83
LK	-0.71	2.1	0	9.6	7.65	7.24	2.77	-16.31	-6.66
CH	-3.76	-4.45	-9.6	0	0.43	-1.54	-9.13	-11.97	-12.5
QRE	-3.79	-4.25	-7.65	-0.43	0	-5.93	-8.83	-10.32	-13.43
QLK	-3.19	-3.35	-7.24	1.54	5.93	0	-8.74	-10.39	-14.3
QCH	-1.04	0.83	-2.77	9.13	8.83	8.74	0	-12.95	-10
RND	1.1	5.93	16.31	11.97	10.32	10.39	12.95	0	-2.9
NI	1.72	4.83	6.66	12.5	13.43	14.3	10	2.9	0

Table 6: Vuong z-scores: MP

References

Carlos Alos-Ferrer and Christoph Kuzmics. Hidden symmetries and focal points. *Journal of Economic Theory*, 148:226–258, 2013.

Ayala Arad and Ariel Rubinstein. The 11–20 money request game: A level-k reasoning study. *The American Economic Review*, 102(7):3561–3573, 2012.

[14] The problem here is actually much simpler than in the very general setting of Vuong (1989).

Sandeep Baliga and Tomas Sjöström. Arms races and negotiations. *The Review of Economic Studies*, 71(2):351–369, 2004.

Volker Benndorf, Ismael Martinez-Martinez, and Hans-Theo Normann. Equilibrium selection with coupled populations in hawk–dove games: Theory and experiment in continuous time. *Journal of Economic Theory*, 165:472–486, 2016.

B Douglas Bernheim. Rationalizable strategic behavior. *Econometrica*, 52:1007–29, 1984.

Sourav Bhattacharya, John Duffy, and SunTak Kim. Voting with endogenous information acquisition: Experimental evidence. *Games and Economic Behavior*, 102:316–338, 2017.

Ken Binmore, Joe Swierzbinski, and Chris Proulx. Does minimax work? An experimental study. *The Economic Journal*, 111(473):445–464, 2001.

Gary E Bolton and Axel Ockenfels. ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, pages 166–193, 2000.

Christoph Brunner, Colin Camerer, and Jacob K Goeree. A correction and re-examination of'stationary concepts for experimental 2x2 games'. *American Economic Review*, 2010.

Terence C Burnham, David Cesarini, Magnus Johannesson, Paul Lichtenstein, and Björn Wallace. Higher cognitive ability is associated with lower entries in a p-beauty contest. *Journal of Economic Behavior & Organization*, 72(1):171–175, 2009.

Colin Camerer, Salvatore Nunnari, and Thomas R Palfrey. Quantal response and nonequilibrium beliefs explain overbidding in maximum-value auctions. *Games and Economic Behavior*, 98:243–263, 2016.

Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004.

Sandra Campo, Emmanuel Guerre, Isabelle Perrigne, and Quang Vuong. Semiparametric estimation of first-price auctions with risk-averse bidders. *The Review of Economic Studies*, 78(1):112–147, 2011.

Kay-Yut Chen and Charles R Plott. Nonlinear behavior in sealed bid first price auctions. *Games and Economic Behavior*, 25(1):34–78, 1998.

Russell W Cooper, Douglas V DeJong, Robert Forsythe, and Thomas W Ross. Selection Criteria in Coordination Games: Some Experimental Results. *American Economic Review*, 80:218–233, 1990.

James C Cox and Ronald L Oaxaca. Is bidding behavior consistent with bidding theory for private value auctions? *Research in Experimental Economics*, 6:131–148, 1996.

Avinash K Dixit and Carl Shapiro. Entry dynamics with mixed strategies. In *The Economics of Strategic Planning*. Lexington Books, Lexington, 1986.

Joseph Farrell. Cheap talk, coordination and entry. *Rand Journal of Economics*, 18:34–39, 1987.

Joseph Farrell and Garth Saloner. Standardization, compatibility, and innovation. *RAND Journal of Economics*, 16(1):70–83, 1985.

Ernst Fehr and Klaus M Schmidt. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, pages 817–868, 1999.

Urs Fischbacher. z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178, 2007.

Drew Fudenberg and Annie Liang. Predicting and understanding initial play. *The American Economic Review*, forthcoming.

Jacob K Goeree and Charles A Holt. Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review*, 91(5):1402–1422, 2001.

Jacob K Goeree and Charles A Holt. A model of noisy introspection. *Games and Economic Behavior*, 46(2):365–382, 2004.

Jacob K Goeree, Charles A Holt, and Thomas R Palfrey. Quantal response equilibrium and overbidding in private-value auctions. *Journal of Economic Theory*, 104(1):247–272, 2002.

Jacob K Goeree, Thomas R Palfrey, and Charles A Holt. Risk averse behavior in generalized matching pennies games. *Games and Economic Behavior*, 45(1):97–113, 2003.

Glenn W Harrison and Elisabet Rutström. Risk aversion in the laboratory. In *Risk Aversion in Experiments*, pages 41–196. Emerald Group Publishing Limited, 2008.

Glenn W Harrison and Elisabet Rutström. Expected utility theory and prospect theory: One wedding and a decent funeral. *Experimental Economics*, 12(2):133, 2009.

John C Harsanyi. Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points. *International Journal of Game Theory*, 2(1):1–23, 1973.

John D Hey and Chris Orme. Investigating generalizations of expected utility theory using experimental data. *Econometrica: Journal of the Econometric Society*, pages 1291–1326, 1994.

Shih-Hsun Hsu, Chen-Ying Huang, and Cheng-Tao Tang. Minimax play at Wimbledon: Comment. *The American Economic Review*, pages 517–523, 2007.

Christoph Kuzmics and Daniel Rodenburger. A case of evolutionarily stable attainable equilibrium in the lab. *Economic Theory*, forthcoming.

Christoph Kuzmics and Brian W Rogers. An incomplete information justification of symmetric equilibrium in symmetric games. SSRN working paper, 2010.

Christoph Kuzmics, Thomas Palfrey, and Brian W Rogers. Symmetric play in repeated allocation games. *Journal of Economic Theory*, 154:25–67, 2014.

John Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, 1982.

John Maynard Smith and George R Price. The logic of animal conflict. *Nature*, 246:15–18, 1973.

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

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

John Nash. Non-cooperative games. *Ph. D. dissertation, Princeton University*, 1950.

Jack Ochs. Games with unique, mixed strategy equilibria: An experimental study. *Games and Economic Behavior*, 10(1):202–217, 1995.

Barry O'Neill. Nonmetric test of the minimax theory of two-person zerosum games. *Proceedings of the National Academy of Sciences*, 84(7):2106–2109, 1987.

Ryan Oprea, Keith Henwood, and Daniel Friedman. Separating the hawks from the doves: Evidence from continuous time laboratory games. *Journal of Economic Theory*, 146(6):2206–2225, 2011.

Ignacio Palacios-Huerta. Professionals play minimax. *The Review of Economic Studies*, 70(2):395–415, 2003.

David G Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica: Journal of the Econometric Society*, 52:1029–51, 1984.

Anatol Rapoport. *Two-person game theory: The essential ideas*. University of Michigan Press, Ann Arbor, 1966.

William H Sandholm. *Population Games and Evolutionary Dynamics*. MIT Press, Cambridge, M.A., 2010.

Reinhard Selten and Thorsten Chmura. Stationary concepts for experimental 2x2-games. *American Economic Review*, 98(3):938–66, 2008.

Reinhard Selten, Thorsten Chmura, and Sebastian J Goerg. Stationary concepts for experimental 2 x 2 games: Reply. *American Economic Review*, 101(2):1041–44, 2011.

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

Dale O Stahl and Paul W Wilson. On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254, 1995.

John B Van Huyck, Raymond C Battalio, and Richard O Beil. Tacit coordination games, strategic uncertainty, and coordination failure. *American Economic Review*, 80:234–248, 1990.

Quang H Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333, 1989.

Mark Walker and John Wooders. Minimax play at Wimbledon. *American Economic Review*, pages 1521–1538, 2001.

Jörgen W Weibull. *Evolutionary Game Theory*. MIT Press, Cambridge, Mass, 1995.

James R Wright and Kevin Leyton-Brown. Beyond equilibrium: Predicting human behaviour in normal form games. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

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

Graz Economics Papers

For full list see:
http://ideas.repec.org/s/grz/wpaper.html
Address: Department of Economics, University of Graz,
Universitätsstraße 15/F4, A-8010 Graz

09–2019 **Philipp Külpmann and Christoph Kuzmics**: On the Predictive Power of Theories of One-Shot Play

08–2019 **Enno Mammen, Jens Perch Nielsen, Michael Scholz and Stefan Sperlich**: Conditional variance forecasts for long-term stock returns

07–2019 **Christoph Kuzmics, Brian W. Rogers and Xiannong Zhang**: Is Ellsberg behavior evidence of ambiguity aversion?

06–2019 **Ioannis Kyriakou, Parastoo Mousavi, Jens Perch Nielsen and Michael Scholz**: Machine Learning for Forecasting Excess Stock Returns The Five-Year-View

05–2019 **Robert J. Hill, Miriam Steurer and Sofie R. Waltl**: Owner Occupied Housing in the CPI and Its Impact On Monetary Policy During Housing Booms

04–2019 **Thomas Aronsson, Olof Johansson-Stenman and Ronald Wendner**: Charity, Status, and Optimal Taxation: Welfarist and Paternalist Approaches

03–2019 **Michael Greinecker and Christoph Kuzmics**: Limit Orders under Knightian Uncertainty

02–2019 **Miriam Steurer and Robert J. Hill**: Metrics for Measuring the Performance of Machine Learning Prediction Models: An Application to the Housing Market

01–2019 **Katja Kalkschmied and Jörn Kleinert**: (Mis)Matches of Institutions: The EU and Varieties of Capitalism

21–2018 **Nathalie Mathieu-Bolh and Ronald Wendner**: We Are What We Eat: Obesity, Income, and Social Comparisons

20–2018 **Nina Knittel, Martin W. Jury, Birgit Bednar-Friedl, Gabriel Bachner and Andrea Steiner**: The implications of climate change on Germanys foreign trade: A global analysis of heat-related labour productivity losses

19–2018 **Yadira Mori-Clement, Stefan Nabernegg and Birgit Bednar-Friedl**: Can preferential trade agreements enhance renewable electricity generation in emerging economies? A model-based policy analysis for Brazil and the European Union

18–2018 **Stefan Borsky and Katja Kalkschmied**: Corruption in Space: A closer look at the world's subnations

17–2018 **Gabriel Bachner, Birgit Bednar-Friedl and Nina Knittel**: How public adaptation to climate change affects the government budget: A model-based analysis for Austria in 2050

16–2018 **Michael Günther, Christoph Kuzmics and Antoine Salomon**: A Note on Renegotiation in Repeated Games [Games Econ. Behav. 1 (1989) 327360]

15–2018 **Meng-Wei Chen, Yu Chen, Zhen-Hua Wu and Ningru Zhao**: Government Intervention, Innovation, and Entrepreneurship

14–2018 **Yu Chen, Shaobin Shi and Yugang Tang**: Valuing the Urban Hukou in China: Evidence from a Regression Discontinuity Design in Housing Price

13–2018 **Stefan Borsky and Christian Unterberger**: Bad Weather and Flight Delays: The Impact of Sudden and Slow Onset Weather Events

12–2018 **David Rietzke and Yu Chen**: Push or Pull? Performance-Pay, Incentives, and Information

11–2018 **Xi Chen, Yu Chen and Xuhu Wan**: Delegated Project Search

10–2018 **Stefan Nabernegg, Birgit Bednar-Friedl, Pablo Muñoz, Michaela Titz and Johanna Vogel**: National policies for global emission reductions: Effectiveness of carbon emission reductions in international supply chains

09–2018 **Jonas Dovern and Hans Manner**: Order Invariant Tests for Proper Calibration of Multivariate Density Forecasts

08–2018 **Ioannis Kyriakou, Parastoo Mousavi, Jens Perch Nielsen and Michael Scholz**: Choice of Benchmark When Forecasting Long-term Stock Returns

07–2018 **Joern Kleinert**: Globalization Effects on the Distribution of Income

06–2018 **Nian Yang, Jun Yang and Yu Chen**: Contracting in a Continuous-Time Model with Three-Sided Moral Hazard and Cost Synergies

05–2018 **Christoph Kuzmics and Daniel Rodenburger**: A case of evolutionary stable attainable equilibrium in the lab

04–2018 **Robert J. Hill, Alicia N. Rambaldi, and Michael Scholz**: Higher Frequency Hedonic Property Price Indices: A State Space Approach

03–2018 **Reza Hajargasht, Robert J. Hill, D. S. Prasada Rao, and Sriram Shankar**: Spatial Chaining in International Comparisons of Prices and Real Incomes

02–2018 **Christoph Zwick**: On the origin of current account deficits in the Euro area periphery: A DSGE perspective

01–2018 **Michael Greinecker and Christopher Kah**: Pairwise stable matching in large economies

15–2017 **Florian Brugger and Jörn Kleinert**: The strong increase of Austrian government debt in the Kreisky era: Austro-Keynesianism or just stubborn forecast errors?

14–2017 **Jakob Mayer, Gabriel Bachner and Karl W. Steininger**: Macroeconomic implications of switching to process-emission-free iron and steel production in Europe

13–2017 **Andreas Darmann, Julia Grundner and Christian Klamler**: Consensus in the 2015 Provincial Parliament Election in Styria, Austria: Voting Rules,Outcomes, and the Condorcet Paradox

12–2017 **Robert J. Hill, Miriam Steurer and Sofie R. Waltl**: Owner Occupied Housing in the CPI and Its Impact On Monetary Policy During Housing Booms and Busts

11–2017 **Philipp Kohlgruber and Christoph Kuzmics**: The distribution of article quality and inefficiencies in the market for scientific journals

10–2017 **Maximilian Goedl**: The Sovereign-Bank Interaction in the Eurozone Crisis

09–2017 **Florian Herold and Christoph Kuzmics**: The evolution of taking roles

08–2017 **Evangelos V. Dioikitopoulos, Stephen J. Turnovsky and Ronald Wendner**: Dynamic Status Effects, Savings, and Income Inequality

07–2017 **Bo Chen, Yu Chen and David Rietzke**: Simple Contracts under Observable and Hidden Actions

06–2017 **Stefan Borsky, Andrea Leiter and Michael Paffermayr**: Product Quality and Sustainability: The Effect of International Environmental Agreements on Bilateral Trade

05–2017 **Yadira Mori Clement and Birgit Bednar-Friedl**: Do Clean Development Mechanism projects generate local employment? Testing for sectoral effects across Brazilian municipalities

04–2017 **Stefan Borsky, Alexej Parchomenko**: Identifying Phosphorus Hot Spots: A spatial analysis of the phosphorus balance as a result of manure application

03–2017 **Yu Chen, Yu Wang, Bonwoo Koo**: Open Source and Competition Strategy Under Network Effects

02–2017 **Florian Brugger**: The Effect of Foreign and Domestic Demand on U.S. Treasury Yields

01–2017 **Yu Chen**: On the Equivalence of Bilateral and Collective Mechanism Design