Economic Theory
https://doi.org/10.1007/s00199-024-01571-y

RESEARCH ARTICLE

The collective wisdom of behavioral game theory

Shu Huang[1] · **Russell Golman**[1]

Received: 18 April 2022 / Accepted: 17 March 2024

Abstract
We apply an algorithm from the wisdom-of-crowds literature to optimally combine behavioral game theory models to more accurately predict strategic choice in one-shot, simultaneous-move games. We find that the optimal weighted average of seven behavioral game theory models predicts out-of-sample choice behavior significantly better than any of the individual models. The crowd of behavioral game theory models is wiser than any single one of them. Different strategic choice models complement each other by capturing distinct patterns of behavior. The field of behavioral game theory is enriched by having this diversity of models.

Keywords Dual accumulator model · Level-k reasoning · Model aggregation · Noisy introspection · Strategic decision making

JEL Classification C72

1 Introduction

When people or businesses must make strategic decisions that interact with the decisions made by other players or stakeholders—for example, setting prices against a rival firm's prices, deciding whether to enter a new market, recruiting and incentivizing workers, or negotiating strategic partnerships—they should use game theory. However, the traditional Nash equilibrium solution concept for simultaneous-move games, based on the assumption of sophisticated, hyper-rational decision makers, poorly predicts actual behavior, especially in novel situations in which players lack experience (i.e., initial play or one-shot games). How then can players anticipate or predict other players' strategic choices?

✉ Russell Golman
rgolman@andrew.cmu.edu

Shu Huang
shuh1@andrew.cmu.edu

[1] Social and Decision Sciences, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA

Published online: 12 April 2024

Behavioral game theory models, based on the psychology of judgment and decision making rather than perfect rationality, predict strategic behavior much more accurately than Nash equilibrium (Camerer 2003). But there are many behavioral models to choose from. The level-k reasoning model (Nagel 1995; Crawford et al. 2013) and the related cognitive hierarchy model (Camerer et al. 2004) rely on the assumption that people only perform a limited number of steps of iterative best-response reasoning. Logit quantal response equilibrium (McKelvey and Palfrey 1995) and the noisy introspection model (Goeree and Holt 1999, 2004) rely on the assumption that people occasionally err and choose suboptimal strategies, but costlier mistakes are less likely. The dual accumulator model (Golman et al. 2020) relies on the assumption that people stochastically sample (or consider) specific scenarios for their counterpart's choice or for their own choice and gradually develop a preference among their own strategies and a belief about their counterpart's anticipated strategy through a linked evidence-accumulation process.[1] Each of these models accounts for stylized facts about actual choice behavior in one-shot simultaneous-move games. With so many models making different predictions, what should a strategic analyst do?

A common approach in much of the behavioral game theory literature is to try to select the model that performs best (Camerer et al. 2004; Wright and Leyton-Brown 2017; Fudenberg and Liang 2019; Golman et al. 2020). However, research on the wisdom of crowds in judgment and decision making and on ensemble methods in machine learning suggests that we may obtain better behavioral predictions by aggregating the predictions of these models instead of trying to select a single one (Dietterich 2000; Davis-Stober et al. 2014). If little were known about the relative performance of the models or their relationships with each other, simply averaging their predictions might make sense. But we can do better.

We use a weighted average to optimally combine the predictions of the set of behavioral game theory models. The optimal weighted average gives more weight to models that are more accurate and also gives less weight to models that are more highly correlated with each other (Winkler 1981; Davis-Stober et al. 2015). Accounting for the models' correlations is particularly important in this context because some of the models are related variants of each other, and a simple average would be distorted by including duplicate predictions, whereas the optimal weighted average aggregates these predictions appropriately. The optimal weighted average based on the correlations between forecasters (along with their variances) performs poorly in some other applications because the weights are highly sensitive to estimation error in the covariance matrix (Clemen and Winkler 1986; Winkler and Clemen 1992), but we use sufficient data for the weighted average to be reliably accurate (Huang et al. 2024).

We use three existing datasets of m individuals playing a set of n one-shot matrix games without feedback (Stahl and Wilson 1995; Külpmann and Kuzmics 2022), which allows us to fit the models at the individual-subject level to account for subject-level heterogeneity. We assess the performance of each model, and of the aggregate weighted average of the models, by computing out-of-sample prediction error using cross validation. Specifically, for each of the individual models, we fit the model to each subject's choices in $n-1$ of the games and then compute the mean squared error

[1] Other-regarding preferences or risk aversion could be added to any of these models as well.

(MSE) of the prediction for the remaining game, and then average this out-of-sample MSE over the m individuals and over the n repetitions with each game being held out. To determine the optimal weights for the weighted average, we fit each model to each subject's choices in $n-2$ of the games and compute the predictions on an $(n-1)^{\text{th}}$ game. Rotating through the out-of-sample predictions for these $n-1$ games, we estimate the covariance matrix of the models' predictions and the covariance between these predictions and the actual choices, and we apply Davis-Stober et al. (2015) formula to obtain the weights. We then use these weights, and the models fit to each subject's choices in the same $n-1$ games, to determine the aggregate prediction for the final held-out game, and again we average the out-of-sample MSE over the m individuals and n repetitions through the held-out games. This process of cross validation allows us to equitably compare the aggregate out-of-sample prediction error to the individual models' out-of-sample prediction errors, without encouraging overfitting, even though the models differ in their flexibility and number of free parameters.

Consistently across the three datasets we find that the optimal weighted average of behavioral game theory models predicts out-of-sample choice behavior significantly better than any of the individual models. This tells us that different models of strategic choice behavior complement each other, by capturing distinct patterns of behavior. Rather than trying to identify the "right" (or best) model, we can make the best predictions by taking advantage of the collective wisdom of this crowd of models.

We gain additional insight by examining the average weights placed on each model (across the n folds of cross validation). The optimal weighted average places about half of its weight on the dual accumulator model, and also gives weight to the level-k reasoning model, a second specification of level-k reasoning with tremble noise, the cognitive hierarchy model, and the noisy introspection model. It gives no weight to the logit quantal response equilibrium model (which is highly correlated with, but not quite as accurate as, the noisy introspection model) or the Nash equilibrium prediction (which predicts quite poorly). Giving the most weight to the dual accumulator model makes sense because it is the most accurate individual model and it does not closely align with any of the other models we consider. Giving weight to the level-k model, the noisy level-k model, the cognitive hierarchy model, and the noisy introspection model helps because they each pick up on other patterns of strategic choice. Specifically, the levels-of-reasoning models may be capturing a particular pattern of limited iterated reasoning, and the noisy introspection model may be capturing an intuitive form of payoff sensitivity.

The approach we take here—leveraging the wisdom of crowds of predictive models—can be applied to make better predictions, and to better understand the diversity of theoretical models, for all kinds of decision making and human behavior. He et al. (2022) use a similar technique to make better predictions about individual choice under risk. This approach could just as well be applied to theories of intertemporal choice, social preferences, or choice under ambiguity, too.

2 Behavioral game theory models

Along with the Nash equilibrium model, we consider six additional models from the behavioral game theory literature that can be applied to simultaneous-move games played once without feedback: two specifications of level-k reasoning (with and without tremble noise), the cognitive hierarchy model, logit quantal response equilibrium, noisy introspection, and the dual accumulator model. These models capture qualitative patterns of strategic choice behavior associated with bounded rationality, i.e., payoff sensitivity (even while best responses remain fixed) and limited iterated reasoning (e.g., in the p-beauty contest game or the traveler's dilemma).[2]

Level-k reasoning proposes that people engage in k steps of best responding, where k varies from person to person, but is typically small. Level-0 individuals are assumed to choose (uniformly) randomly. Level-k individuals then choose best responses to the level-$(k-1)$ choice, for $k > 0$. (If multiple best responses exist, we assume that choice is uniformly random among them.) We fit the parameter k to individual subject's choices. In our specification with tremble noise, the prediction becomes a convex combination of the level-k best response (with weight $(1-\epsilon)$ and the uniform mixed strategy (with weight ϵ), where the weight ϵ is also an individual-level free parameter.

The cognitive hierarchy model resembles level-k reasoning in that individuals engage in a small number of steps of best responding, but in this model, individuals best respond to the mixture of strategies from everybody who engages in fewer steps of reasoning. The model assumes that the number of steps of best responding that people perform follows a Poisson distribution (when determining the mixed strategy that each individual responds to).

In the logit quantal response equilibrium, an individual with "rationality" parameter λ plays a mixed strategy ω_1 while expecting his counterpart to play a mixed strategy ω_2, such that the probability of strategy s_i for the focal player is

$$\omega_1(s_i) = \frac{e^{\lambda u(s_i;\,\omega_2)}}{\sum_k e^{\lambda u(s_k;\,\omega_2)}}$$

and the belief about the probability of strategy s_j for the counterpart is

$$\omega_2(s_j) = \frac{e^{\lambda u(s_j;\,\omega_1)}}{\sum_k e^{\lambda u(s_k;\,\omega_1)}}.$$

That is, ω_1 and ω_2 are logit responses to each other. The rationality parameter governs how noisy the logit response is—if $\lambda = 0$, it produces a uniform mixed strategy, whereas for $\lambda \to \infty$, it approaches a perfect best response. This parameter, λ, can be fit at the subject level.

[2] Cooperative behavior in the prisoner's dilemma should be attributed to social preferences, not bounded rationality. Models of social preferences can be evaluated in the absence of strategic uncertainty (e.g., in the dictator game), and could then be integrated with any of the models of strategic thinking that we consider here.

The noisy introspection model predicts a mixed strategy

$$\omega = \lim_{n\to\infty} \phi_\mu(\phi_{\tau\mu}(\ldots, \phi_{\tau^n\mu}(\omega_0))),$$

where ϕ_μ is the logit response function with rationality parameter $\lambda = \frac{1}{\mu}$ and ω_0 is the uniform mixed strategy. The "error" parameter μ determines a baseline noisiness of responses, and the "telescope" parameter $\tau > 1$ determines how increasingly noisy each successive higher-order belief is. Both of these parameters can be fit at the individual subject level.

The dual accumulator model assumes that an individual alternately samples one of his counterpart's strategies or one of his own strategies to accumulate activation for his own strategies or his counterpart's strategies, respectively. The probability of sampling the counterpart's strategy s_j at time step t is

$$p_j = \frac{e^{\lambda A_{2j}(t-1)}}{\sum_k e^{\lambda A_{2k}(t-1)}},$$

and if strategy s_j is sampled, the activation for his own strategies becomes $A_{1i}(t) = A_{1i}(t-1) + u(s_i;\, s_j)$ for each strategy s_i. Similarly, the probability of sampling his own strategy s_i at time step t is

$$p_i = \frac{e^{\lambda A_{1i}(t)}}{\sum_k e^{\lambda A_{1k}(t)}},$$

and if strategy s_i is sampled, the activation for his counterpart's strategies becomes $A_{2j}(t) = A_{2j}(t-1) + u(s_j;\, s_i)$ for each strategy s_j. Activation for all strategies is initially zero and then accumulates for T steps of sampling, at which point the strategy with the highest activation is chosen. (If multiple strategies equally achieve maximum activation at time step T, then choice is uniformly random among them.) The activation for the counterpart's strategies does not directly influence the choice (given the activation for one's own strategies), but co-evolves with the sampling probabilities, which then affect the activation for one's own strategies. The "stochastic sampling" parameter λ and the "time limit" parameter T are fit at the individual subject level.

Omitted models

We exclude models incorporating social preferences, such as team reasoning (Bacharach 1999), as well as combinations of the included models with social-preference models of altruism, reciprocity, fairness, or inequality-averse preferences (Levine 1998; Fehr and Schmidt 1999; Bolton and Ockenfels 2000; Charness and Rabin 2002; Falk and Fischbacher 2006). Incorporating social preferences would no doubt improve the accuracy of all of our models. Nevertheless, social preferences can be explored in more parsimonious settings that do not require players to engage with complex strategic considerations (see Charness and Rabin 2002). For simplicity, we

focus exclusively on models of strategic thinking here. We exclude models incorporating risk- and ambiguity-preferences (Goeree et al. 2003; Eichberger and Kelsey 2011; Beggs 2021). For the same reason, we exclude set-valued solution concepts as well (Goeree et al. 2005; Goeree and Louis 2021; Barberà et al. 2022). We also exclude models intended to describe learning in games or long-run behavior after learning has occurred, such as experience-weighted attraction (Camerer and Hua Ho 1999), instance-based learning (Gonzalez et al. 2003), action-sampling equilibrium (Selten and Chmura 2008), payoff-sampling equilibrium (Osborne and Rubinstein 1998), and impulse-balance equilibrium (Selten and Chmura 2008). Repeated games that allow for learning should be considered in future work, but for simplicity, we focus here on one-shot interactions, in which players do not have the opportunity to learn. We hope that a better understanding of initial play will inform models of learning and repeated play.

3 Predictive accuracy

3.1 Data

We use three existing data sets reporting the behavior of individuals playing sets of one-shot games without feedback. The first data set, reported by Stahl and Wilson (1995), consists of the strategy choices of 48 subjects who each played a set of 12 3×3 symmetric games, including three games with unique mixed-strategy Nash equilibria and nine games with unique pure-strategy symmetric equilibria, some (but not all) of which being dominance solvable. The second data set, reported by Külpmann and Kuzmics (2022), consists of the strategy choices of 147 subjects who each played a set of 20 2×2 games of the hawk-dove and matching-pennies forms. The third data set, also reported by Külpmann and Kuzmics (2022), consists of the strategy choices of 166 subjects who each played a set of 20 3×3 games, half of them of a hawk-dove variety with an additional strategy available and half of them of a rock-paper-scissors variety.[3] In each dataset we know each individual subject's choice in each game, so we can fit the models at the individual-subject level. The Stahl and Wilson (1995) games all have payoffs between 0 and 100, and the Külpmann and Kuzmics (2022) games all have payoffs between 0 and 10, so we need not worry about normalizing payoffs differently across games in the same dataset.

3.2 Out-of-sample mean squared error

We use out-of-sample mean squared error (MSE) to evaluate the predictive accuracy of each model. For each game played by each subject, the MSE is $\frac{1}{C}\sum_{i=1}^{C}(x_i - y_i)^2$, where C is the number of strategies in each game, x_i is the predicted probability that the individual selects strategy s_i, and y_i is an indicator variable for whether the individual actually played strategy s_i. To determine the out-of-sample prediction that a model makes for a particular subject playing a particular game, we first fit that model

[3] All of the game payoffs are laid out in the original papers.

to that subject's choices in the other $n-1$ games in that dataset by minimizing the mean squared error averaged over these games. Additional details about our model fitting procedure are provided in the SM appendix. For each model, we then average the out-of-sample MSEs across all n games and m subjects in that dataset.

4 Model aggregation

We consider a weighted average of the seven models described above, with a 7×1 vector of non-negative weights $\mathbf{w}$, such that $\mathbb{1}^T\mathbf{w} = 1$, i.e., the weights sum to one. If the weights were set with indicator variables for the most accurate model, then the weighted average would simply agree with the most predictive model. Alternatively, if the weights were all equal, $\mathbf{w} = \frac{1}{7}\mathbb{1}$, then the weighted average would simply be the simple average, which works surprisingly well to extract the wisdom of a crowd of forecasters. Davis-Stober et al. (2015) provide a system of equations to determine the optimal weights that minimize MSE, given forecasters' variances and correlations (with each other and with the target variable). The optimal weighted average $\mathbf{w}$ (in the case of unbiased forecasters) is the solution to

$$\begin{bmatrix} \Sigma_{XX} & \mathbb{1} \\ \mathbb{1}^T & 0 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{w} \\ \lambda \end{bmatrix} = \begin{bmatrix} \sigma_{Xy} \\ 1 \end{bmatrix} \tag{1}$$

with $w_\iota \geq 0$ for all ι, where Σ_{XX} is the covariance matrix for the forecasters, σ_{Xy} is the vector of covariances between each forecaster and the target variable, and λ is a real-valued unknown variable, i.e., a Lagrange multiplier (Davis-Stober et al. 2015). To apply this formula to aggregate our models, we need to estimate the covariance matrix for the models' predictions.

For each game we determine the weighted average of the models by looking only at the models' predictions (and the actual data) from the other $n-1$ games, to ensure we are always comparing out-of-sample predictions. But within those $n-1$ games, we still seek out-of-sample predictions from each of the models to use to estimate the covariance matrix for the models. So, we perform another hold-one-out analysis within this sample of $n-1$ games. For each of these $n-1$ games, we fit each model to each subject's choices in $n-2$ of the games and then determine the model's prediction for the $(n-1)^{\text{th}}$ game. We then compute the sample covariance matrix for these out-of-sample predictions. We plug these values into Eq. 1 to determine the weights on each of the seven models. And we then generate the aggregate-model prediction for each player in the held-out n^{th} game by using this weighted average to combine the predictions of the seven models, where, as before, each model's parameters are determined by fitting the models to that subject's choices in the other $n-1$ games. Thus, the aggregate model makes out-of-sample predictions for each player in each game. (Accordingly, the precise weights used in the aggregate model may vary from game to game.) The overall MSE of the aggregate model is the MSE averaged across all n games and m subjects in that dataset.

5 Results

5.1 Correlation structure of the individual models

Before we assess the performance of the weighted average relative to the individual models, we first consider the relationships between the individual models. The weights on the individual models depend not only on the accuracy of these models (i.e., the size of their out-of-sample MSEs for the games used to determine the optimal weights), but also on the correlations between these models (i.e., the extent to which their errors tend to be in the same direction as other models' errors). Figure 1 presents the average correlation coefficients between every pair of individual models (across all folds of cross validation) for each dataset. Naturally, the two specifications of level-k reasoning and the cognitive hierarchy model are all highly correlated with each other. The logit quantal response equilibrium and noisy introspection models are highly correlated with each other in the two Külpmann and Kuzmics (2022) datasets as well. Some of the correlations vary considerably between the datasets because each experiment presented subjects with different games.

5.2 Accuracy of the individual models and the weighted average

Figure 2 shows the average out-of-sample MSE for each individual behavioral game theory model (black circles) and for the optimal weighted average of these models (red square) for each dataset. The data points shown in gray display the average out-of-sample MSEs across the subjects for each of the games distinctly. Figure 3 displays box plots showing the distribution of out-of-sample MSEs at the subject-level (averaged across the games) for each dataset. In these figures we see that the optimal weighted average has consistently lower out-of-sample MSEs than any of the individual models. (The specific p-values from paired t-tests are reported in Tables B1-B6 in the Appendix, with $p < .05$ for all comparisons, except the comparison between the weighted average and the dual accumulator model for MSEs across games in the Külpmann and Kuzmics (2022) 2×2 games dataset and the comparison between the weighted average and the dual accumulator model for MSEs across subjects in the Stahl and Wilson (1995) dataset.) The crowd of behavioral game theory models is wiser than any single one of them. Having multiple models is not a problem forcing us to make a difficult choice among them. Rather, the collection of models gives us a way to describe a variety of patterns of strategic choice behavior and to make better predictions.

5.3 Weights on the individual models

Figure 4 shows box plots of the weights placed on each model (across the n folds of cross validation) when using the optimal weighted average. The dual accumulator model gets about half of the weight (ranging from $M = 46\%$, averaging across the cross validation, in the Külpmann and Kuzmics (2022) 3×3 games to $M = 74\%$ in the Külpmann and Kuzmics (2022) 2×2 games). It gets the most weight because

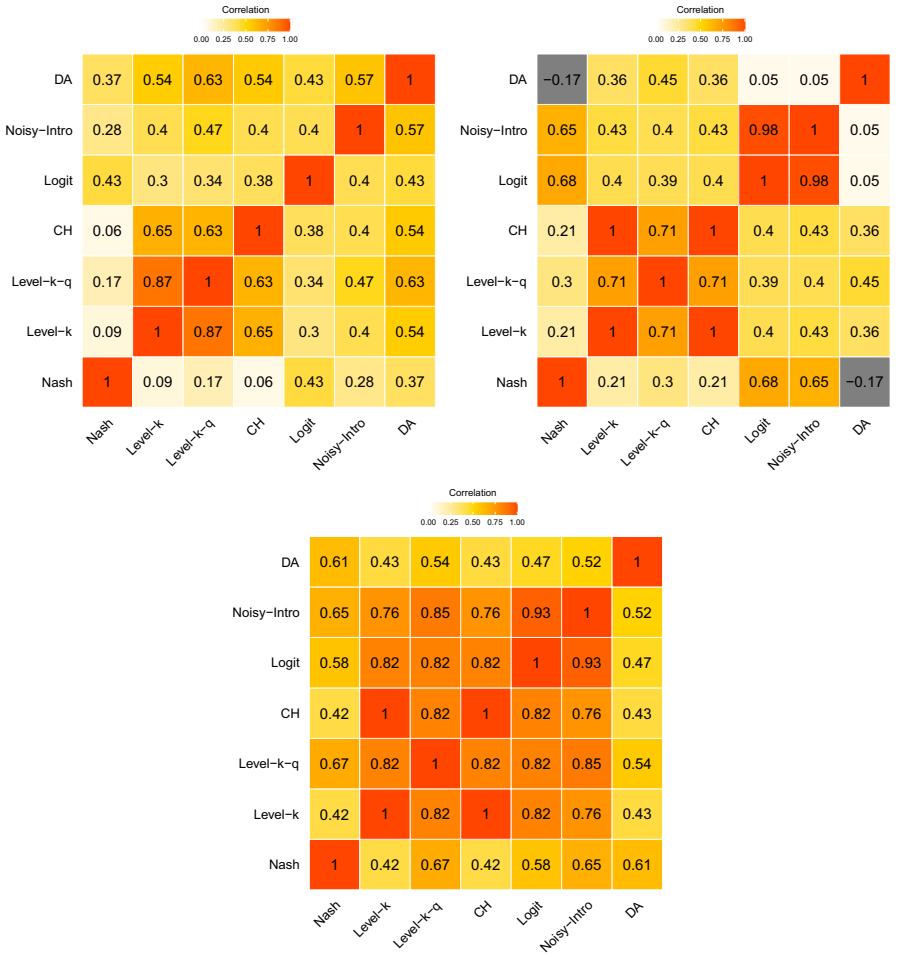

Fig. 1 Average correlation coefficients between pairs of behavioral game theory models. Top left: Stahl and Wilson (1995) 3 × 3 games. Top right: Külpmann and Kuzmics (2022) 2 × 2 games. Bottom: Külpmann and Kuzmics (2022) 3 × 3 games (color figure online)

it is the most accurate individual model and it does not closely align with any of the other models we consider. The optimal weighted average also places weight on the the cognitive hierarchy model, the level-k reasoning model (for two of the datasets), the noisy level-k reasoning model (for two of the datasets), and the noisy introspection model (for two of the datasets). It gives no weight to the logit equilibrium model and the Nash equilibrium model. The weights (or lack thereof) on these models cannot be viewed in isolation. In both Külpmann and Kuzmics (2022) datasets the level-k reasoning model and the cognitive hierarchy model always make the same predictions, so they share weight equally; but in the Stahl and Wilson (1995) dataset the level-k reasoning model is similar to, but not quite as accurate as, the cognitive hierarchy model and the noisy level-k model, so the pure level-k model does not get any weight

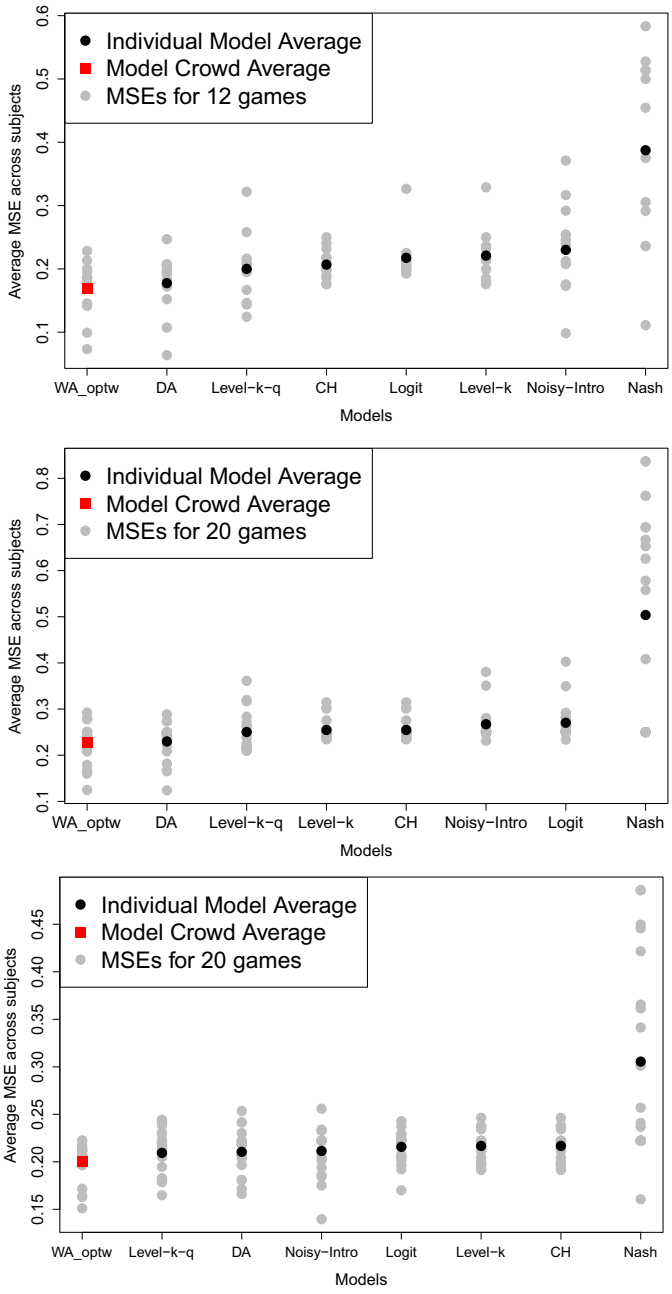

Fig. 2 Out-of-sample MSEs averaged across subjects (gray data points corresponding to each distinct game) for each individual behavioral game theory model and for the optimal weighted average. Top: Stahl and Wilson (1995) 3×3 games. Middle: Külpmann and Kuzmics (2022) 2×2 games. Bottom: Külpmann and Kuzmics (2022) 3×3 games (color figure online)

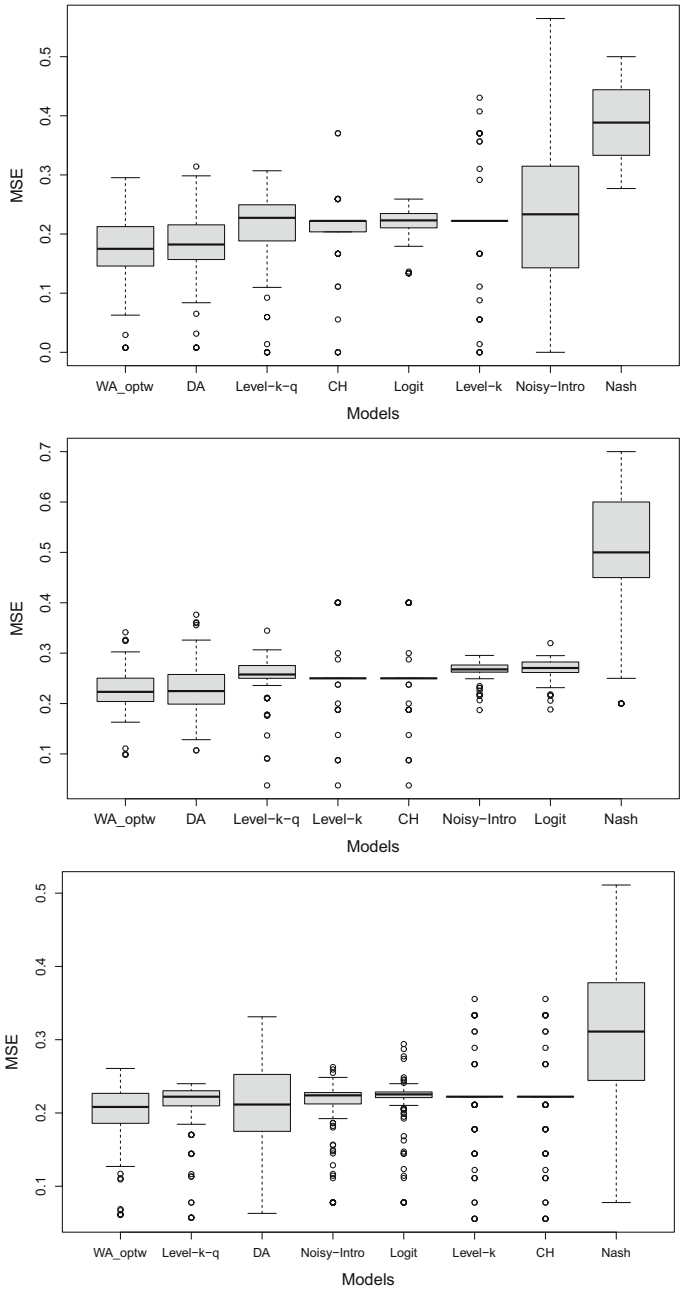

Fig. 3 Box plots showing the distribution of out-of-sample MSEs at the subject-level (averaged across games) for each individual behavioral game theory model and for the optimal weighted average. Top: Stahl and Wilson (1995) 3×3 games. Middle: Külpmann and Kuzmics (2022) 2×2 games. Bottom: Külpmann and Kuzmics (2022) 3×3 games

there. If the cognitive hierarchy model and the noisy level-k model were omitted from the analysis, then the pure level-k model would get positive weight for this dataset as well. The substantial weight placed on the noisy introspection model for the Stahl and Wilson (1995) dataset is notable because the model appears to suffer from overfitting in this setting. On its own, it actually performs worse in out-of-sample prediction for this dataset than the complete ignorance model that always predicts a pure uniform distribution. While this overfitting leads to poorly calibrated predictions and thus poor performance on its own, the model still provides a valid signal and adds value when combined with the other models.

The weights given to the individual models are fairly stable across the folds of the cross validations (as shown in Fig. 4). Tables B10-B12 in the Appendix report the weights for each fold of the cross validation. The robustness of the weighted average can also be seen in the game-by-game MSEs reported in Tables B7-B9. While the games vary in their predictability, the weighted average performs consistently well in comparison to the individual models.

We also see in Tables B7-B9 that the weighted average outperforms a simple average of the models, which does not account for the accuracy of the individual models or the correlations between them. One advantage of the optimal weighted average, relative to the simple average of the models, is that the weighted average recognizes the poor performance of the Nash equilibrium prediction and consequently gives it no weight. Another advantage of the weighted average is that its performance is more robust to the inclusion of additional models that are similar to other already-included models. Including related models like the cognitive hierarchy and noisy level-k models along with the pure level-k model effectively double or triple counts this class of models in the simple average, but the weighted average adjusts the weighting to maintain robust overall performance. While the weights placed on the individual models are highly sensitive to the set of models included in the analysis, the performance of the weighted average is actually less sensitive to the choice of which models to include. Other weighting methods, such as optimal weights under the assumption that errors are unbiased and uncorrelated with true values (Lamberson and Page 2012), the contribution weighted model (Budescu and Chen 2014), or stacking multiple weighting methods (Huang et al. 2023), also adjust weights depending on the set of individual models included in the analysis, and may perform similarly well or even better. Our results demonstrate that there is room for the combination of behavioral game theory models to outperform any single one of them.

6 Conclusion

We have applied an algorithm from the *wisdom-of-crowds* literature to the field of *behavioral game theory* to more accurately predict strategic choice in one-shot, simultaneous-move games. By optimally combining the predictions of various behavioral game theory models, we can make better predictions than by using any single model alone. The success of the aggregate prediction shows us that these different

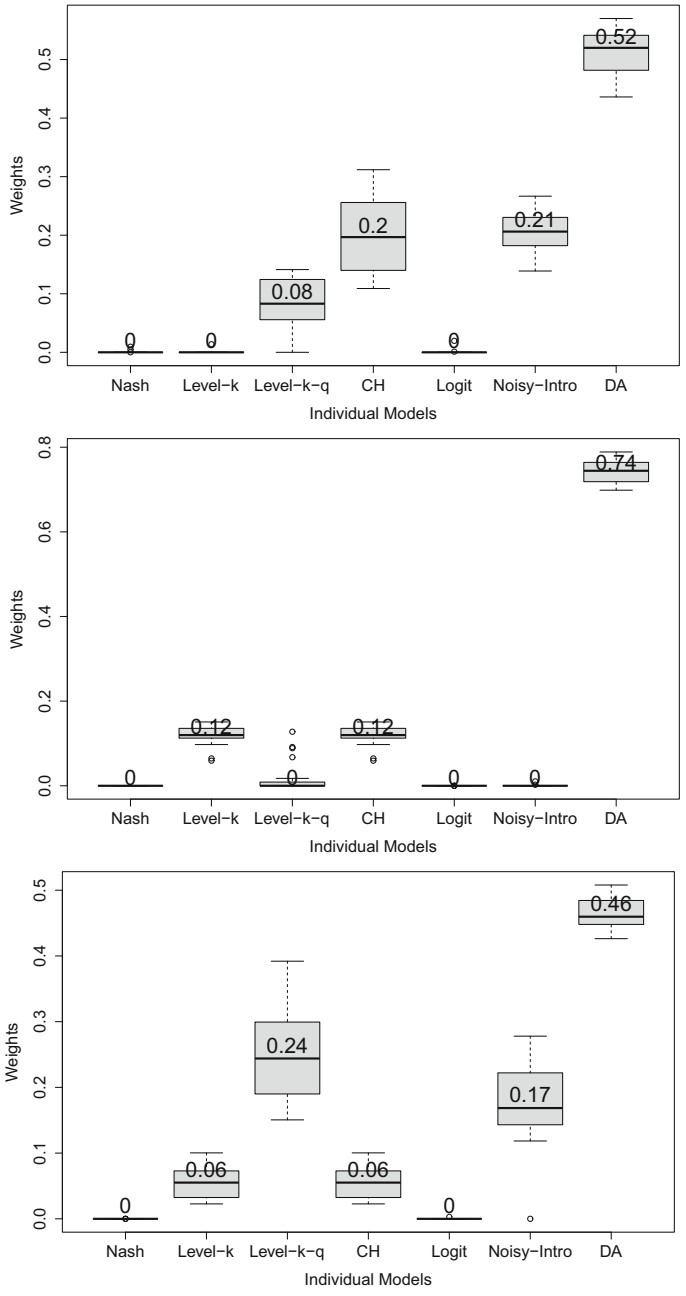

Fig. 4 Weights placed on each behavioral game theory model when using the optimal weighting method. Box plots display the variation over the n folds of cross validation. Median weights reported here. Top: Stahl and Wilson (1995) 3×3 games. Middle: Külpmann and Kuzmics (2022) 2×2 games. Bottom: Külpmann and Kuzmics (2022) 3×3 games

models are capturing distinct aspects of real, human strategic decision making. The field of behavioral game theory is enriched by having this diversity of models.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00199-024-01571-y.

Funding Open Access funding provided by Carnegie Mellon University.

Declarations

Conflict of interest We have no Conflict of interest to declare that are relevant to the content of this article. The data and code to run our analyses is publicly available at https://osf.io/a47gs/?view_only=ba91691b4c054253b3c85143a67f7eec

Appendix A Model fitting

The Nash equilibrium model has no free parameters to fit, but does not always generate a unique prediction. We select among the set of Nash equilibria by tracing the principal branch of the logit quantal response equilibrium from the uniformly mixed strategy to the limiting logit equilibrium, which is a standard (Nash) equilibrium selection method (McKelvey and Palfrey 1995).[4]

For most of the behavioral models (all except the cognitive hierarchy model), we minimize MSE between predicted and observed choices by searching over a uniform grid of parameter values. For the cognitive hierarchy models, we iteratively estimate the individual levels of reasoning and the distribution of levels in the population until they converge.

For the specifications of level-k reasoning and the cognitive hierarchy model, we search $k \in \{0, ..6\}$. For the specification of level-k with noise, we also search $\epsilon \in \{0, .01, \ldots, 1\}$. For the logit quantal response equilibrium we search $\lambda \in \{0.1, 0.2, \ldots, 10\}$. (We find the logit equilibrium by repeatedly computing logit responses starting from a uniform mixed strategy, and stopping when the average absolute value of the difference in probabilities from one step to the next is less than .01. This procedure selects one primary equilibrium in games with multiple logit equilibria.) For the noisy introspection model, we search $\mu \in \{.001, .002, \ldots, .01, .02, \ldots, 10\}$ and $\tau \in \{1, 1.1, \ldots, 10, 11, \ldots, 100, 110, \ldots, 1000\}$. (Here too we compute logit responses until the average absolute value of the difference in probabilities from one step to the next is less than .01.) For the dual accumulator model, we search

[4] The only games in our datasets with multiple Nash equilibria are the hawk-dove games, where this method effectively selects the symmetric Nash equilibrium of each game.

$\lambda \in \{0, .01, \ldots, 0.1, 0.2, \ldots, 10\}$ and $T \in \{1, 2, \ldots, 50\}$ when fitting the Stahl and Wilson (1995) dataset and search $\lambda \in \{2^{-10}, 2^{-9}, \ldots, 2^{10}\}$ and $T \in \{1, 2, \ldots, 30\}$ when fitting the Külpmann and Kuzmics (2022) datasets. (We computed the predicted choice probabilities by simulating 10, 000 runs of the model.)

References

Bacharach, M.: Interactive team reasoning: A contribution to the theory of co-operation. Res. Econ. **53**(2), 117–147 (1999)

Barberà, S., De Clippel, G., Neme, A., Rozen, K.: Order-k rationality. Econ. Theor. **73**(4), 1135–1153 (2022)

Beggs, A.: Games with second-order expected utility. Games Econom. Behav. **130**, 569–590 (2021)

Bolton, G.E., Ockenfels, A.: ERC: A theory of equity, reciprocity, and competition. Am. Econ. Rev. **91**(1), 166–193 (2000)

Budescu, D.V., Chen, E.: Identifying expertise to extract the wisdom of crowds. Manag. Sci. **61**(2), 267–280 (2014)

Camerer, C., Hua Ho, T.: Experience-weighted attraction learning in normal form games. Econometrica **67**(4), 827–874 (1999)

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

Camerer, C.F., Ho, T.H., Chong, J.K.: A cognitive hierarchy model of games. Q. J. Econ. **119**(3), 861–898 (2004)

Charness, G., Rabin, M.: Understanding social preferences with simple tests. Q. J. Econ. **117**(3), 817–869 (2002)

Clemen, R.T., Winkler, R.L.: Combining economic forecasts. J. Bus. Econ. Stat. **4**(1), 39–46 (1986)

Crawford, V.P., Costa-Gomes, M.A., Iriberri, N.: Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. J. Econ. Lit. **51**(1), 5–62 (2013)

Davis-Stober, C.P., Budescu, D.V., Broomell, S.B., Dana, J.: The composition of optimally wise crowds. Decis. Anal. **12**(3), 130–143 (2015)

Davis-Stober, C.P., Budescu, D.V., Dana, J., Broomell, S.B.: When is a crowd wise? Decision **1**(2), 79 (2014)

Dietterich, T.G.: Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Systems, pp. 1–15. Springer (2000)

Eichberger, J., Kelsey, D.: Are the treasures of game theory ambiguous? Econ. Theor. **48**(2–3), 313–339 (2011)

Falk, A., Fischbacher, U.: A theory of reciprocity. Games Econom. Behav. **54**(2), 293–315 (2006)

Fehr, E., Schmidt, K.M.: A theory of fairness, competition, and cooperation. Q. J. Econ. **114**(3), 817–868 (1999)

Fudenberg, D., Liang, A.: Predicting and understanding initial play. Am. Econ. Rev. **109**(12), 4112–41 (2019)

Goeree, J.K., Holt, C.A.: Stochastic game theory: For playing games, not just for doing theory. Proc. Natl. Acad. Sci. **96**(19), 10564–10567 (1999)

Goeree, J.K., Holt, C.A.: A model of noisy introspection. Games Econom. Behav. **46**(2), 365–382 (2004)

Goeree, J.K., Holt, C.A., Palfrey, T.R.: Risk averse behavior in generalized matching pennies games. Games Econom. Behav. **45**(1), 97–113 (2003)

Goeree, J.K., Holt, C.A., Palfrey, T.R.: Regular quantal response equilibrium. Exp. Econ. **8**, 347–367 (2005)

Goeree, J.K., Louis, P.: M equilibrium: A theory of beliefs and choices in games. Am. Econ. Rev. **111**(12), 4002–4045 (2021)

Golman, R., Bhatia, S., Kane, P.B.: The dual accumulator model of strategic deliberation and decision making. Psychol. Rev. **127**(4), 477–504 (2020)

Gonzalez, C., Lerch, J.F., Lebiere, C.: Instance-based learning in dynamic decision making. Cogn. Sci. **27**(4), 591–635 (2003)

He, L., Analytis, P.P., Bhatia, S.: The wisdom of model crowds. Manag. Sci. **68**(5), 3635–3659 (2022)

Huang, S., Broomell, S.B., Golman, R.: A hypothesis test algorithm for determining when weighting individual judgments reliably improves collective accuracy or just adds noise. Decision **11**(1), 7–34 (2024)
Huang, S., Golman, R., Broomell, S.B.: Combining the aggregated forecasts: An efficient method for improving accuracy by stacking multiple weighting models (2023) (Working paper)
Külpmann, P., Kuzmics, C.: Comparing theories of one-shot play out of treatment. J. Econ. Theory **205**, 105554 (2022)
Lamberson, P., Page, S.E.: Optimal forecasting groups. Manag. Sci. **58**(4), 805–810 (2012)
Levine, D.K.: Modeling altruism and spitefulness in experiments. Rev. Econ. Dyn. **1**(3), 593–622 (1998)
McKelvey, R.D., Palfrey, T.R.: Quantal response equilibria for normal form games. Games Econom. Behav. **10**(1), 6–38 (1995)
Nagel, R.: Unraveling in guessing games: An experimental study. Am. Econ. Rev. **85**(5), 1313–1326 (1995)
Osborne, M.J., Rubinstein, A.: Games with procedurally rational players. Am. Econ. Rev. **88**, 834–847 (1998)
Selten, R., Chmura, T.: Stationary concepts for experimental 2x2-games. Am. Econ. Rev. **98**(3), 938–966 (2008)
Stahl, D.O., Wilson, P.W.: On players models of other players: Theory and experimental evidence. Games Econom. Behav. **10**(1), 218–254 (1995)
Winkler, R.L.: Combining probability distributions from dependent information sources. Manag. Sci. **27**(4), 479–488 (1981)
Winkler, R.L., Clemen, R.T.: Sensitivity of weights in combining forecasts. Oper. Res. **40**(3), 609–614 (1992)
Wright, J.R., Leyton-Brown, K.: Predicting human behavior in unrepeated, simultaneous-move games. Games Econom. Behav. **106**, 16–37 (2017)