# Model comparisons using tournaments: Likes, "dislikes", and challenges

Leonidas Spiliopoulos and Andreas Ortmann
Australian School of Business, University of New South Wales

# Online supplemental appendices

## Appendix B
## Modeling game and role-conditional heterogeneity in heuristic use

Our attention will focus on analyzing the tournament data using the 7s-model due to its performance in the tournament. Our analysis is also informed by our "submissions" for the tournament, the conclusions of which are interesting. We examine whether heuristics use is conditional on game classes and on the player's role in a game (as a first- or second-mover). Finally, we compare how well our models can predict human behavior for each game class and role.

Let $g \in G$ denote the game class, $r \in R = \{Player\,1, Player\,2\}$ denote the role of a subject, and $A_r$ denote the action choice sets for a player in role $r$, where $A_1 = \{In, Out\}$ and $A_2 = \{Left, Right\}$. The structure of the dataset made available by EER consists of a set of observations, indexed by $i$; each observation is comprised of the proportion of subjects who chose the *In* (or *Right*) action for a given game and role $p\,(a_i)$, and the remaining variables are the payoffs of the game (which uniquely determine the game that is being played) and the role of the subjects within that game. Therefore, each observation corresponds to subjects' behavior for a unique combination of a specific game (note, not game class) and role within that game. Let the estimated coefficients for each of the seven strategies $s \in S$, as presented in Table A3, be given by $\beta_s^{r,g}$.

$$\begin{aligned} p(a_i) &= \beta_{rl}^{r,g} Rtnl_i + \beta_{nr}^{r,g} Nrtnl_i + \beta_{mxmn}^{r,g} Mxmn_i + \beta_{L1}^{r,g} L1_i + \beta_{jmx}^{r,g} Jmx_i + \beta_{Mxwk}^{r,g} Mxwk_i + \beta_{mndff}^{r,g} Mndff_i \\ & \sum_{r,s,g} \beta_s^{r,g} = 1 \end{aligned} \tag{1}$$

Since the model does not have a constant and the sum of all coefficients is constrained to equal one, the estimated coefficients can be interpreted as the proportion use of a strategy in the subject pool. The objective function in Eq. 1 is the RMSD of the observed and predicted probabilities of the choices given the chosen training dataset $\Omega$, which is a subset of the whole dataset.

$$RMSD = \sqrt{\min_{\hat{\beta}_s^{r,g}} \frac{1}{\mid \Omega \mid} \sum_{\omega \in \Omega} \left[p(a_\omega) - \hat{p}(a_\omega)\right]^2} \tag{2}$$

The model presented above permits different heuristic use according to game class and role. The assumption of game and role conditional heterogeneity can be tested empirically by comparing the performance of the unrestricted model to that of three nested models created by imposing the following restrictions on coefficients:

1. Subjects use heuristics in the same proportions regardless of the game class $\forall g \in G : \beta_s^{r,g} = \beta_s^r$.

2. Subjects use heuristics in the same proportions regardless of their role in the game $\forall r \in R : \beta_s^{r,g} = \beta_s^g$.

3. Subjects use heuristics in the same proportions regardless of the game class and their role $\forall r \in R, \forall g \in G : \beta_s^{r,g} = \beta_s$.

The robustness of the estimation procedure for these models with respect to data partitioning will be examined using different cross-validation regimes. Let the chosen training and cross-validation datasets be denoted as $\Omega$ and $\Omega'$ respectively, then $\Omega \rightarrow \Omega'$ is a specific regime. The first two regimes involve training the model on the tournament's estimation dataset, and cross-validating on the tournament's prediction dataset ($est \rightarrow pred$) and vice-versa, using the *pred* dataset as the training dataset and the *est* dataset as the cross-validation dataset ($pred \rightarrow est$)—this is 2-fold cross-validation. For a specific regime, the cross-validation performance criterion $CV_{\Omega \rightarrow \Omega'}$ is given by Eq. 3 where $\hat{p}(a_{\omega'_k})$ is evaluated at the values of $\hat{\beta}_s^{r,g}$ estimated from the set $\Omega$—ideally, the results from the two regimes ($est \rightarrow pred$) and ($pred \rightarrow est$) should be very similar.

$$CV_{\Omega \rightarrow \Omega'} = \sqrt{\sum_{\omega'_k \in \Omega'_k} \left[ p(a_{\omega'_k}) - \hat{p}(a_{\omega'_k}) \right]^2} \tag{3}$$

Another regime uses the combined data from the estimation and prediction datasets, without distinction, and performs leave-one-out cross-validation (referred to as *l1o*), which is a special case of $k$-fold cross-validation. For general $k$-fold cross-validation, let $\Omega'_k$ denote the cross-validation set for the $k$th fold. The cross-validation performance criterion averaged over all folds, to take into account all possible data-partitioning decisions, is given by Eq. 4 where $\hat{p}(a_{\omega'_k})$ is evaluated at the values of $\hat{\beta}_s^{r,g}$ estimated from the set $\Omega_k$.

$$CV = \sqrt{\frac{1}{k} \sum_k \frac{1}{\mid \Omega'_k \mid} \sum_{\omega'_k \in \Omega'_k} \left[ p(a_{\omega'_k}) - \hat{p}(a_{\omega'_k}) \right]^2} \tag{4}$$

Specifically in the case of *l1o*, a single observation is excluded from the training sample and the model is estimated on the remaining 239 games—performance is estimated on the test set comprised of the single observation. This is repeated for each of 240 possible partitions of the dataset into training and cross-validation datasets; therefore, the cross-validation criterion in this case is created by averaging over all the possible cross-validation sets $k = 240$.

In order to estimate these models, there are some additional assumptions that need to be made regarding special cases where heuristics are perfectly correlated or the sample size of games for a given game class is too small to reliably estimate coefficients.

The models estimated for each player role do not use the full range of the defined strategies for the following reasons: *Nrtnl* does not make a prediction for Player 1, whilst for player 2 the *Rtnl, Mxmn* and *L1* strategies are perfectly correlated. The affected coefficients are automatically dropped from the estimation procedure.

Although some games belong to more than one class, in our estimation procedures we assign games to one class using the following rules. If two classes overlap, but neither is a subset of the other, then we add the intersection of the game classes as a new game class, for example, we include a game class *ssci,* which is the overlap of *ss* and *ci*. If one game class is a complete subset of another class (or union of classes), we assign games as belonging to the subsumed set whenever possible (e.g., trust games), which are a subset of costly help, are assigned to the *tr* class and not the *ch* class.

Sometimes the number of free parameters are too few compared to the number of observations within each game classes to efficiently estimate a model permitting game conditional heterogeneity. For robustness, we solve this problem in two different ways; however, since significant differences are not found, the results of the first solution are presented in the main discussion in Appendix B—results of the second solution are presented for comparison in Tables B4 and B5.

1. The game classes with less than twenty observations, *fp, rp, fh, ch* and *tg,* are pooled into a new single class, denoted by *po*. Admittedly, this is arbitrary, as there is no particular reason to believe that these game classes are related to justify this categorization. However, this approach does not throw any of the data away as the next approach does—this leads to seven games classes, $G = \{ci, nd, ss, cp, po, ssci, nt\}$.

2. Simply drop the *fp, rp, fh, ch* and *tg* game classes from the estimation procedure and use only the remaining game classes, $G = \{ci, nd, ss, cp, ssci, nt\}$—we refer to this as the *reduced* dataset.

Finally, a remaining issue when estimating a model with imposed player role homogeneity is the fact that *Nrtnl* does not make any prediction for player 1. We assume that players using the *Nrtnl* strategy as a second player would use the *Rtnl* strategy when moving first. Note, that although some of the heuristics were perfectly correlated for player 2 choices, this is not the case for player 1 choices; therefore, when estimating jointly both player roles these heuristics are no longer perfectly correlated.

**Results**

The estimation results for the models with various restrictions on heterogeneity and combinations of estimation and prediction sets are provided in Table B1. Within each of the three cross-validation regimes, the best performing model assumes no game class heterogeneity but permits role heterogeneity. However, the maximum difference in the cross-validation RMSD in all these cases compared to a model with game and role homogeneity is only equal to 0.001—roughly one additional correct prediction per one thousand. This is neither economically significant nor statistically significant at the 5% level, as determined by a paired Wilcoxon signed rank ($z = 0.508, p = 0.61$) and a sign test ($p = 0.37$) on the CV for each observation in the *l1o* procedure. We conclude that subjects in the experiment did not significantly condition their heuristic use on the game class or on their role in the game. This result is surprising in light of the existing literature that finds significant strategic adaptation (Payne, Bettman, & Johnson, 1988, 1993; Rieskamp, 2008; Rieskamp & Hoffrage, 1999)—this discrepancy is likely the result of the implementation details of the EER tournament, discussed in Section *Implementation and the Duhem-Quine problem.*

Comparing the two cross-validation regimes *est* $\rightarrow$ *pred* and *pred* $\rightarrow$ *est*, will shed light on the robustness of tournament results in light of the random sampling of games and subjects. The cross-validation RMSD for all subjects and roles are clearly smaller in the *est* $\rightarrow$ *pred* procedure compared to *pred* $\rightarrow$ *est* for all combinations of game and role heterogeneity. Closer inspection of this result by disaggregating the CV by player role reveals that this large difference can be attributed to the player 1 role. Comparing the model with no game or role heterogeneity across the two procedures reveals a moderately large difference in CV for player 1 (0.018), but a small difference for player 2 (0.003). We believe that such a procedure should be carried out for any tournament as it provides an indication of the magnitude of the random effects occurring due to sampling (either game or subject sampling)—differences in models should be interpreted relative to these sampling errors. For example, for player 1, the difference between the best performing model and the 7s-model is 0.006 a value that is three times smaller than the induced change in performance by switching the cross-validation regime. We conclude that the submitted models are

not significantly different from the 7s-model. The difference between the best performing player 2 model and the baseline is 0.005 compared to an induced change of 0.003—this results seems more robust to the partitioning of the data into training and cross-validation sets.

Further examination of the robustness of the cross-validation procedures can be performed by comparing the parameter estimates presented in Table B2. Instability in the estimates for *Rtnl* and *Nrtnl* is revealed, however this is primarily due to the high degree of collinearity of these two heuristics, and not specifically due to the differences in the cross-validation procedures. Table B3 presents the correlation between the actions prescribed by the various heuristics. The heuristics *Rtnl, Nrtnl, L1* and *Mxmn* exhibit very high correlation for the games examined in the tournament, therefore inference with respect to which of these heuristics best represents subjects' behavior is limited. Further data would be required to distinguish between these two heuristics effectively. For the remaining parameter estimates, the largest difference occurs for *L1*, with the two procedures yielding a difference in 4.5 percentage points—again the effects seem to be of moderate size.

Table B2 presents the estimated proportions of heuristic use for the model with no game class or role heterogeneity. The most widely used heuristics are *Rtnl* (and *Nrtnl* for the second player), followed by *Mxmn* and *L1*. These heuristics are not strongly correlated with the remaining heuristics incorporating social preferences, permitting stronger conclusions to be drawn from such a comparison. As noted earlier, heuristics employing social preferences are used by a significantly smaller proportion of subjects than heuristics only incorporating own payoffs. Note that the use of *Nrtnl*, implies that that subjects did care about their opponents' payoffs but only if their own payoffs would not be affected negatively.

Table B1
*Cross-validation performance (root-mean-square deviation) for various model assumptions*

| | *l1o* | | | | *est* → *pred* | | | | *pred* → *est* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Role het.? | No | | Yes | | No | | Yes | | No | | Yes | |
| Game het.? | No | Yes | No | Yes | No | Yes | No | Yes | No | Yes | No | Yes |
| Player 1 | 0.100 | 0.105 | 0.100 | 0.105 | 0.090 | 0.095 | 0.090 | 0.115 | 0.108 | 0.116 | 0.108 | 0.118 |
| Player 2 | 0.061 | 0.064 | 0.060 | 0.067 | 0.066 | 0.071 | 0.064 | 0.071 | 0.063 | 0.069 | 0.062 | 0.076 |
| Mean | 0.083 | 0.087 | 0.083 | 0.088 | 0.079 | 0.084 | 0.078 | 0.096 | 0.089 | 0.095 | 0.088 | 0.099 |

Table B2
*Estimated percentage use of strategies (%)*

| Dataset | *Rtnl* | *Nrtnl* | *L1* | *Mxmn* | *Jmx* | *Mxwk* | *Mndff* |
|---|---|---|---|---|---|---|---|
| *l1o* | 19.6 | 26.8 | 17.7 | 19.7 | 6.7 | 6.1 | 3.3 |
| *est* → *pred* | 11.2 | 33.2 | 20.1 | 18.9 | 6.9 | 6.6 | 3.0 |
| *pred* → *est* | 35.7 | 12.6 | 15.6 | 20.3 | 6.4 | 5.8 | 3.7 |

Table B3
*Correlation of heuristic decisions*

| | *Rtnl* | *Nrtnl* | *L1* | *Mxmn* | *Jmx* | *Mxwk* | *Mndff* |
|---|---|---|---|---|---|---|---|
| *Rtnl* | 1.000 | 0.982 | 0.754 | 0.783 | 0.365 | 0.412 | 0.067 |
| *Nrtnl* | 0.982 | 1.000 | 0.741 | 0.772 | 0.392 | 0.425 | 0.069 |
| *L1* | 0.754 | 0.741 | 1.000 | 0.826 | 0.393 | 0.399 | -0.015 |
| *Mxmn* | 0.783 | 0.772 | 0.826 | 1.000 | 0.322 | 0.360 | 0.052 |
| *Jmx* | 0.365 | 0.392 | 0.393 | 0.322 | 1.000 | 0.650 | -0.014 |
| *Mxwk* | 0.412 | 0.425 | 0.399 | 0.360 | 0.650 | 1.000 | 0.316 |
| *Mndff* | 0.067 | 0.069 | -0.015 | 0.052 | -0.014 | 0.316 | 1.000 |

Table B4
*Cross-validation performance (root-mean-square deviation) for various model assumptions for the reduced dataset*

| | *l1o* | | | | *est* → *pred* | | | | *pred* → *est* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Role het.? | No | | Yes | | No | | Yes | | No | | Yes | |
| Game het.? | No | Yes | No | Yes | No | Yes | No | Yes | No | Yes | No | Yes |
| Player 1 | 0.095 | 0.098 | 0.095 | 0.097 | 0.084 | 0.083 | 0.085 | 0.098 | 0.102 | 0.104 | 0.102 | 0.103 |
| Player 2 | 0.059 | 0.059 | 0.058 | 0.063 | 0.063 | 0.069 | 0.060 | 0.062 | 0.053 | 0.063 | 0.053 | 0.072 |
| Mean | 0.079 | 0.081 | 0.079 | 0.082 | 0.074 | 0.076 | 0.073 | 0.082 | 0.082 | 0.086 | 0.081 | 0.089 |

Table B5
*Estimated proportions of strategy use for the reduced dataset*

| Dataset | *Rtnl* | *Nrtnl* | *L1* | *Mxmn* | *Jmx* | *Mxwk* | *Mndff* |
|---|---|---|---|---|---|---|---|
| *l1o* | 15.9 | 29.9 | 18.8 | 19.1 | 6.4 | 6.6 | 3.2 |
| *est* → *pred* | 14.6 | 30.0 | 19.8 | 18.4 | 6.9 | 7.5 | 2.7 |
| *pred* → *est* | 19.2 | 28.0 | 18.3 | 19.4 | 5.6 | 5.8 | 3.8 |

**Predictive power of the model per game class and role.** A comparison of the predictive power of the model with no game or role heterogeneity, as measured by cross-validation performance per game class and player role is presented in Table B6. Note, a single model is estimated on all the data for all game classes and player roles; however, we report the disaggregated performance per game class and role to allow for more informative comparisons. These results, including confidence intervals, were also discussed briefly and presented graphically in Figure 2 in the main text.

As expected, for all game classes except *nd*, the cross-validation RMSD for player 1 is higher than for player 2—this is due to the uncertainty that the first player faces with respect to player 2's move, which adds to the difficulty in modeling behavior. The model's performance in predicting behavior in each game class, aggregated over player roles, is (ordered from best performance to worst): *ssci, ss, rp, nd, nt, ci, fh, tg, cp* and *fp*. The games *fp* and *cp* are not surprisingly the most difficult to predict as they require beliefs about the likelihood of opponents punishing their behavior (when it is not rational). The standard seven strategies do not incorporate beliefs about punishment—this may be one of the limitations of the 7-s model and may have contributed to

the poor performance for these games. Evidence that this explanation is valid are the tournament results for player 1, where the top four models are based on the standard seven strategies with modifications accounting for fear of punishment and beliefs about player 2.

Table B6
*Cross-validation (root-mean-square deviation) prediction accuracy per game class*

| Game class | *ci* | *nd* | *ss* | *cp* | *fp* | *rp* | *fh* | *ch* | *tg* | *ssci* | *nt* | All classes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Player 1 | 0.123 | 0.075 | 0.073 | 0.159 | 0.220 | 0.077 | 0.124 | 0.085 | 0.125 | 0.053 | 0.093 | 0.100 |
| Player 2 | 0.042 | 0.079 | 0.060 | 0.046 | 0.164 | 0.053 | 0.042 | 0.073 | 0.051 | 0.038 | 0.066 | 0.061 |
| Both players | 0.092 | 0.077 | 0.067 | 0.117 | 0.194 | 0.066 | 0.092 | 0.079 | 0.095 | 0.046 | 0.081 | 0.083 |

*Note.* Common-interest=*ci*, Near-dictator=*nd*, Safe-shot=*ss*, Strategic-dummy=*sd*, Costly-punish=*cp*, Free-punish=*fp*, Rational-punish=*rp*, Free-help=*fh*, Costly-help=*ch*, Trust-game=*tg*, Safe-shot/common-interest=*ssci*, Other-non-trivial=*nt*.

## Appendix C
## Robustness to environmental variation

Let an (sampling) environment, $e$, be uniquely defined by the probability distribution over the set of game classes, $G$—this is the probability that a player would face a game of this class. We examine the eleven game classes discussed thus far, therefore the universal set of environments $E$ is a standard 11-dimensional simplex, where $\pi_g^e$ is the probability a game of class $g$ is drawn in a specific environment $e$: $E = \{(\pi_1, ..., \pi_{11}) \in \mathbb{R}^{11} \mid \sum_G \pi_g = 1, \pi_g \geq 0\}$.

We adopt a semi-Bayesian approach in dealing with our uncertainty regarding what the true environmental is. Ideally, the researcher would impose a prior over the universe of possible environments (the distribution of game classes or specific payoff sampling schemes) and run the same experiment for each environment. The posterior probability that a model is the best performer, accounting for the possible environments, is then easily calculated. Clearly, running the same experiment many times for each possible environment is practically impossible. However, we propose the following solution that only requires a single experiment using a unique sampling scheme. Environmental variation can be approximated by simply re-weighting the importance of observations (conditional on their game class) in the model estimation procedures. Let $\pi_g$ be the proportion of games belonging to class $g$ in a tournament and $\pi_g^e$ be the desired proportion of the target environment denoted by $e$—the weights per game class are given by $\pi_g^e/\pi_g$. Models can be estimated by the following re-weighted objective function, Eq. 5, where $w_\omega$ is the weight for each observation in the training set determined by the game class the observation belongs to—cross-validation criteria are similarly adjusted.[1] To simultaneously control for sampling variation, the cross-validation technique we employ here is the leave-one-out procedure discussed above.

$$RMSD = \sqrt{\min_{\hat{\beta}_s^{r,g}} \frac{1}{\mid \Omega \mid} \sum_{\omega \in \Omega} w_\omega \left[p(a_\omega) - \hat{p}(a_\omega)\right]^2} \tag{5}$$

[1] An alternative approach, which could also account for sampling variation, would be to create new datasets by repetitively sampling with replacement from the set of experimental games with probabilities defined by the environment $e$. However, this would increase the computational cost to impractical levels.

**Analysis**

We perform the robustness analysis by re-estimating the 7s-model and the Charness-Rabin (CR) model—the two best performing baseline models according to ENO—and also examine the subgame perfect Nash equilibrium. The set of environments, $E$, that we will examine will consist of all possible combinations of the eleven game classes with restrictions to manage the computation costs by discretizing the linear combinations of game classes and bounding the weights:[2] $\pi_g^e \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}, \sum_{g \in G} \pi_g^e = 1$. We impose a uniform prior distribution over this set of 92,378 environments, $p(e) = \mid E \mid^{-1}$. Our environmental variation criterion, $R$, is simply the probability that a model is the best performer over the set of environments. If $R$ values are significantly different from 1, this implies that model performance depends strongly on the type of environment. Therefore, further examination of the relationship between environments and individual model performance is warranted; we believe that this is invaluable information for guiding the development of future models.

We proceed in our analysis by using the data from both the estimation and prediction datasets and the leave-one-out cross validation procedure discussed above. The subgame perfect Nash equilibrium was outperformed in every single environment therefore is not mentioned further. Define the representative environment of a model as the average game class distribution over the environments where this model was the best performer—this permits a comparison of when specific models perform well. Table C1 presents the result of the robustness analysis. The 7s-model outperforms the CR model in 92% of the environments examined for the first player. Examining the representative environments it is very clear that the CR performs better when the proportion of free punish games is high—note, this confirms the evidence presented in Figure 2 that the 7s-model exhibits its worst performance at predicting behavior for this class of games. The mean difference in RMSD between the 7s-model and the CR model over all environments is 0.019; therefore, the 7s-model will make roughly two more correct predictions out of every one hundred. For the second player, the 7s-model outperforms the CR model in all the environments examined, and the mean difference in RMSD is equal to 0.0204.

The superiority of the 7s-model over the Charness-Rabin model and subgame-perfect Nash equilibrium is confirmed for a very large space of possible environments, and is not due to the specific sampling environment used in the tournament (although this result is of course dependent on the auxiliary assumptions and implementation details of the tournament). With respect to whether the top performing submissions would also be significantly different from each other for a large proportion of the environment space we remain skeptical. As we do not have the code for these submissions we are unable to perform these calculations. However, given that the Charness-Rabin and 7s-models had significantly larger differences in performance for the tournament's sampling scheme than the differences in the top performing models (as presented in Table A5), it is likely that this will lead to non-robust differences over the entire environment space.

[2]The computational cost of performing this analysis is very large and quickly gets out of control—each environment requires the estimation of a model 240 times (due to leave-one-out cross-validation) for a total of 22,170,720 optimization problems per player role.

Table C1
*Robustness to environmental variation*

| Player | Model | $R$ | Representative environment | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *ci* | *nd* | *ss* | *cp* | *fp* | *rp* | *fh* | *ch* | *tg* | *ssci* | *nt* |
| 1 | 7s | 0.92 | 0.09 | 0.09 | 0.09 | 0.09 | 0.08 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 |
| 1 | CR | 0.08 | 0.08 | 0.08 | 0.07 | 0.06 | 0.22 | 0.08 | 0.08 | 0.08 | 0.09 | 0.08 | 0.07 |
| 2 | 7s | 1 | - | - | - | - | - | - | - | - | - | - | - |
| 2 | CR | 0 | - | - | - | - | - | - | - | - | - | - | - |

References

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 534–552.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*. Cambridge University Press.

Rieskamp, J. (2008). The importance of learning when making inferences. *Judgment and Decision Making*, *3*(3), 261–277.

Rieskamp, J., & Hoffrage, U. (1999). When do people use simple heuristics, and how can we tell? In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 141–167). New York, NY, US: Oxford University Press.