Profiling Players in Dynamic Games: A Mobile Experiment*

Konrad Grabiszewski[†] Alex Horenstein [‡]

Abstract

This article has two objectives. First: explore what mobile technology can offer for experimental research by way of creating *Blues and Reds*, a mobile app designed to conduct experiments on dynamic game theory. Second: design a method of profiling players in dynamic games and test its predictive power using data from the app. A two-dimensional profile depicts a subject's quality and speed of reasoning. With 35,826 observations from 6,463 subjects located in 141 countries, we replicate the same test of predictive power in 22 different games and confirm that a subject's profile predicts whether she behaves consistently with backward induction.

JEL Codes: C72, C73, C99.

Key words: dynamic games, profiling, experimental game theory, mobile experiment.

*We are grateful to Andrew Leone, who as the Vice Dean for Faculty Development and Research at the University of Miami Business School, helped us secure financing to conduct this project. For their comments, suggestions, and feedback, we thank Raphael Boleslavsky, Christopher Cotton, Kfir Eliaz, Simon Grant, John Hey, Ariel Rubinstein, and Emanuel Vespa as well as the participants of our presentations at conferences (2017 ASU Ph.D. Reunion, 2018 North American Econometric Society, 2018 Asian Econometric Society, 2018 Tsinghua BEAT, 2018 North American ESA, 2018 Spring Midwest Economic Theory, 2018 D-TEA, 2018 FUR, 2018 WIEM, 2019 Asia-Pacific ESA) and seminars (ITAM, Keio University, Kindai University, Kansai University, KAUST, University of Miami, MBSC, American University of Armenia).

[†]Mohammad bin Salman College, Saudi Arabia; konrad.grabiszewski@gmail.com

[‡]Department of Economics, University of Miami, FL, USA; horenstein@bus.miami.edu

I Introduction

In the past decade, mobile technology has become a global phenomena present in every corner of the world. The omnipresence of mobile technology opens attractive, yet unexplored, opportunities to conduct large-scale experiments. To take advantage of these opportunities, we employed a team of developers to create *Blues and Reds*, a mobile app for iOS and Android devices.[1] The app's objective is to run experiments – due to their nature called "mobile experiments" – on dynamic game theory. Everyone with access to Google Play or the App Store can become a subject in a mobile experiment and there are billions of people with such access.

This article uses data from *Blues and Reds* and advances a novel method of profiling players in dynamic games that satisfies two natural requirements suggested by Rubinstein (2016): descriptive and predictive powers. The subject's profile mirrors the reasoning process in dynamic games (descriptive power) and predicts whether the subject behaves consistently with backward induction (predictive power). What increases the reliability of the predictive power of the proposed profiling is successful replication. The same qualitative results are replicated in 22 various dynamic games, where each game is treated as a separate experiment, played by thousands of subjects from all over the world.

Each of the 22 games is a finite dynamic game with perfect and complete information which are played by a human subject against Artificial Intelligence (AI). The subject either wins or loses; there are no ties. AI is designed to be fully rational and to exploit the subject's mistakes. Games vary in the number of actions (from 2 to 4) and rounds (from 3 to 6).

In every game, there is only one set of actions which lead to the subject winning the game. This set constitutes a unique equilibrium path. If, in any round, the subject makes a mistake and chooses an action that is not consistent with backward induction, the subject loses the game. Not making a mistake is the fundamental measure of the subject's performance in a game.

Blues and Reds records the following data for each subject and each game: (i) whether a

[1] The team consisted of software and database engineers, graphic and animation designers, art consultant, music composer, and testers. *Blues and Reds* is the authors' creation in the sense of the idea: namely, what the app and experiment should be about and its appearance. However, in the actual development, the authors' roles were more of producers rather than app developers. *Blues and Reds* has been available in four languages (English, Spanish, Chinese traditional, and Chinese simplified) for free since August 2017. In March 2019, *Blues and Reds* had 4.1 (out of 5) stars in Google Play ranking, placing it among very popular titles produced by multi-billion-dollar companies; e.g., *Candy Crush Saga* (4.4 stars, Activision Blizzard), *Tetris* (4.0 stars, Electronic Arts), and *Super Mario Run* (3.8 stars, Nintendo).

subject wins, and (ii) for every round, the subject's response time. Response time at a given round is the time measured in seconds that a subject spends on deciding what action to choose. Recording round-by-round response times makes the data from *Blues and Reds* unique and permits for a more refined analysis of behavior.

The proposed profiling is two-dimensional since two questions need to be answered to depict a reasoning process in dynamic games: "how" people think and "how much" they think. The former refers to the approach of finding a solution, and the latter is about the effort of employing that approach. The objective is to measure "how" and "how much" a subject reason by looking at her response times.

Since the selection of strategy takes place at the very beginning of a game, the "how" dimension is measured as percentage of the relative response time at the first round ($RRT1$) which is the total response time allocated to the first round; $RRT1 = \frac{RT1}{TT} \times 100\%$ where $RT1$ is the response time at the first round and TT is the total response time (sum of RTs from all rounds). For a savvy subject, that percentage is relatively higher compared to a naive one. The "how much" dimension is the total time spent on reasoning in a game (TT). Fast thinkers have lower total time compared to slower subjects. Four profiles emerge: savvy-fast, savvy-slow, naive-fast, and naive-slow.

Prior to testing the predictive power of the suggested profiling, it is necessary to hypothesize which profile would be ranked higher in the sense of the likelihood of behaving in accordance with backward induction. The ranking is not necessarily obvious as profiles are vectors rather than scalars. Given that solving games is a cognitive task, we argue that lexicographic ranking is a natural ordering of profiles. Therefore, Ann is a higher profile than Bob if she is either savvier than him or, assuming they are equally savvy, she is faster.

Data analyses from *Blues and Reds* shows that higher profiles are, indeed, more likely to choose consistently with backward induction. Figure I summarizes the findings of this article.

[Figure I about here.]

This article relates to different streams of literature in economics. Due to its data-collection method, it is important to mention innovative methodologies like newspaper-based experiments (e.g., Bosch-Domènech et al. (2002)) and online experiments (e.g., Ariel Rubinstein's `gametheory.tau.ac.il`, Chen and Konstan (2015), Chen et al. (2014), and Liu et al. (2014)). One of several advantages of mobile experiments are the ease and low-cost of engaging large groups of people as subjects; centralized promotion tools like Google AdWords make the task particularly effective and efficient.

Methodologically, this article belongs to the experimental literature that relies on measuring response times. In economics, this literature started with Rubinstein (2006) and, since then, has grown very fast (e.g., Rubinstein (2007), Piovesan and Wengstr ́om (2009), Rand et al. (2012), Rubinstein (2013), Schotter and Trevino (2014), Agranov et al. (2015), Evans et al. (2015), Clithero (2016), Rubinstein (2016), Gill and Prowse (2017), Lohse et al. (2017), and Spiliopoulos and Ortmann (2017)). The differentiating factor presented in this article is that the data includes round-by-round response times rather than just the coarser total response time.

The findings in this article contribute to the literature which profiles players in games: the level-k model (e.g., Stahl and Wilson (1994), Stahl and Wilson (1995), and Nagel (1995), Ho et al. (1998), Costa-Gomes et al. (2001), Bosch-Domènech et al. (2002), Costa-Gomes and Crawford (2006), Costa-Gomes and Weizsäcker (2008), Wang et al. (2010), Agranov et al. (2012), Arad and Rubinstein (2012), Ho and Su (2013), Burchardi and Penczynski (2014), Hargreaves Heap et al. (2014), Shapiro et al. (2014), Georganas et al. (2015), Fehr and Huck (2016), Penczynski (2016), and Batzilis et al. (2017)), the cognitive hierarchy model (Camerer et al. (2004)), or studies that measure subjects' cognitive skills (e.g., Burks et al. (2009), Burnham et al. (2009), Palacios-Huerta and Volij (2009), Rydval et al. (2009), Agranov et al. (2012), Brañas-Garza et al. (2012), Carpenter et al. (2013), Duffy and Smith (2014), Agranov et al. (2015), Alaoui and Penta (2016), Allred et al. (2016), Bayer and Renou (2016), Benito-Ostolaza et al. (2016), Fehr and Huck (2016), Gill and Prowse (2016), Hanaki et al. (2016), and Kiss et al. (2016)). Since the second dimension of the proposed profiling relates to the speed of thinking, the closest article related to this paper are Rubinstein (2013), Rubinstein (2016), and the concept of fast/slow thinking in Kahneman (2013). The unique features of this article are its emphases on dynamic games and a two-dimensional structure of the proposed profiling.

The reminder of this article proceeds as follows. Section II provides a detailed description of *Blues and Reds* as an experiment. Given that it is freely available, readers are encouraged to download *Blues and Reds* to experience the experiment. (Links to download *Blues and Reds* from Google Play and the App Store are on the website `www.bluesandreds.com`.) Section III presents the construction of profiles and examines its predictive power. Section IV concludes.

II Data

II.A *Blues and Reds* as an experiment

The first 4 games in *Blues and Reds* constitute the mandatory practical tutorial. This article uses data from 22 games played after completion of the tutorial collected between August 15, 2017 to February 6, 2018.

Each of the 22 games resembles a game-theoretic tree; Figure II depicts an example. A game starts with the subject choosing which blue bridge the RoboToken (golden sphere) crosses. Then AI selects the red bridge for the RoboToken. And so on: subject chooses at odd rounds, and AI at even rounds. If the RoboToken ends at a blue node, the subject wins; otherwise, the subject loses.

[Figure II about here.]

Winning in *Blues and Reds* requires the subject to follow the path that is the same as the unique winning path selected by the backward induction algorithm. Deviating from that path results with the subject losing. Hence, winning is indicative of the subject backward inducting.

Games have a symmetrical structure: the number of actions at each node of a given round is the same. A 3-round game is denoted as $N1.N2.N3$ where Ni is the number of actions at the ith round. A 4-round game is labeled as $N1.N2.N3.N4$, and so on. Figure II depict the 3.2.2.2 game. The first column in Table I in Section II.B includes the list of all games from the dataset.

Subjects can play each game only once — there is no second chance if they lose. Finally, the sequence in which the games appear to the subjects is randomized for each subject.

II.B Data description

For each game and each subject, *Blues and Reds* records whether a subject wins or loses and the time (measured in seconds) a subject spends on selecting actions at each round. From the data, the following three variables are constructed:

1. TT. This is subject's total response time in a game; i.e., the sum of round-based RTs.

2. $RRT1$. This is a subject's relative time spent at the first round defined as $\frac{RT1}{TT} \times 100\%$ where $RT1$ is the subject's response time in the first round of a game.

3. Win. This variable takes the value 1 if subject wins and is otherwise 0.

Data cleaning was approached conservatively and observations with a total time above the 95th percentile within each game were removed from the sample. The final data consists of 35,826 observations generated from 6,463 subjects located in 141 countries. Figure III depicts the geographical distribution of subjects.

[Figure III about here.]

For each game in the experiment, Table I presents the number of subjects, percentage of subjects who won (i.e., behave in accordance with backward induction) and summary statistics for $RT1$ and $RRT1$.

[Table I about here.]

III Profiles in Dynamic Games

Consider four fictional subjects who played the game 3.2.2.2 (Figure II) and whose response times as well as total times are presented in Table II. Subjects only choose at odd rounds; hence, the data consists only of RT1 and RT3.

[Table II about here.]

Given the data in Table II, the following challenges are of interest. First, designing a subject's profile that depicts the reasoning process in dynamic games. Second, proposing a ranking of profiles that orders subjects from the most to the least likely to behave in accordance with backward induction. Third, testing whether that ranking indeed predicts rational behavior.

III.A Profiles: describing reasoning in dynamic games

Solving any problem – not just finding an optimal strategy in a dynamic game – is a two-dimensional process consisting of "how" to reason and "how much" to reason. Measuring "how" and "how much" people reason depends on the specific task at hand. In the context of dynamic games, the proposed metrics are based on the subject's response times and the fact that the first round of a game is the crucial thinking time for selecting an optimal strategy. This is especially true in *Blues and Reds* as making a mistake in any round results with a loss. If a correct strategy was selected in round 1, then, during the following rounds, a subject only spends time on physically picking the right actions but no longer has to re-think what to choose.

Consequently, the first dimension – "how" a subject reasons – is depicted by $RRT1$, the relative time a subject spends in the first round ($RRT1 = \frac{RT1}{TT} \times 100\%$). Allocating time to later rounds (i.e., lower $RRT1$) indicates a possible flaw in a subject's reasoning (e.g., guessing or making a mistake in the first round). High $RRT1$ is consistent with correctly following the backward induction algorithm. Subjects with high $RRT1$ are called savvy; those with low $RRT1$ are labeled naive. In Table II, Ann and Bob are equally savvy ($RRT1 = 0.75$) and savvier than Chris and David who are also equally savvy ($RRT1 = 0.4$).

The second dimension – "how much" a subject reasons – is, simply, captured by TT, the subject's total response time. Fast subjects are those with low TT, while slow are characterized by high TT. In terms of their thinking speed, in Table II, Ann and Chris are fast, while Bob and David are slow.

III.B Profiles: lexicographic ranking

The next challenge lies in comparing Ann and Bob with profiles ($RRT1_A$, TT_A) and ($RRT1_B$, TT_B), respectively. This comparison is necessary to establish the predictive power of the proposed profiling. After all, the goal is to empirically verify that higher-ranked profiles are more likely to behave in line with backward induction. However, this test demands a definition of what it means when Ann is ranked higher than Bob.

With scalar profiles, there is not much to discuss regarding ranking methods. For profiles like ($RRT1$, TT), ranking is not a trivial task as there are several ways to rank vectors. Lexicographic ranking seems to be a natural solution.

Solving games is a cognitive task. If a person does not understand how to find the optimal strategy, additional time will not help her with the task. Hence, how people reason ($RRT1$) is more important than how much they reason (TT). Consequently, a savvy profile is higher than a naive profile, no matter how fast/slow both are.

However, the second dimension (TT) becomes useful for ranking people who are equally knowledgeable about solving games (the same $RRT1$). Those more proficient or experienced with choosing strategies reason faster (a lower TT) as thinking is cognitively costly and spending less time on reasoning is preferable. For instance, an expert in game theory and her student share the same understanding of how to apply the backward induction algorithm. However, the expert requires less time to implement the algorithm as she has more experience solving games. Consequently, assuming two profiles are equally savvy, faster one is a higher profile.

To summarize, Ann with ($RRT1_A$, TT_A) is said to have a higher profile than Bob with ($RRT1_B$, TT_B) if one of the following holds.

1. $RRT1_A > RRT1_B$.
2. If $RRT1_A = RRT1_B$, then $TT_A < TT_B$.

Ann is the highest profile in Table II, followed by Bob, then Chris, and David as the lowest profile.

III.C Profiles: testing predicting power

In this Section, the same test of predictive power is replicated in 22 various games. Each time, the following exercise is conducted. First, subjects are divided into savvy and naive according to their $RRT1$; savvy are those with $RRT1$ above the median $RRT1$. Second, for each $RRT1$-group, subjects are divided into fast and slow; fast are those with TT below the median TT. Therefore, for each game, there are four groups of subjects according to the proposed two-dimensional profiling: savvy-fast, savvy-slow, naive-fast, and naive-slow.

Data is presented in 22 tables, with one table for each game. Each table has the same structure which resembles Figure I from the Introduction and is explained in a sample Table III. For each profile $i = \{savvy - fast, savvy - slow, naive - fast, naive - slow\}$, the percentage of subjects of profile i who won (i.e., behaved in accordance with backward induction) is calculated and denoted as P_i. The main hypothesis is that, for each game, the following three inequalities hold.

$$P_{savvy-fast} > P_{savvy-slow} \tag{1}$$

$$P_{savvy-slow} > P_{naive-fast} \tag{2}$$

$$P_{naive-fast} > P_{naive-slow} \tag{3}$$

A t-statistic is provided at the bottom of a table for each inequality.

[Table III about here.]

[Table IV about here.]

[Table V about here.]

[Table VI about here.]

[Table VII about here.]

[Table VIII about here.]

[Table IX about here.]

[Table X about here.]

[Table XI about here.]

[Table XII about here.]

[Table XIII about here.]

[Table XIV about here.]

[Table XV about here.]

[Table XVI about here.]

[Table XVII about here.]

[Table XVIII about here.]

[Table XIX about here.]

[Table XX about here.]

[Table XXI about here.]

[Table XXII about here.]

[Table XXIII about here.]

[Table XXIV about here.]

[Table XXV about here.]

Analyses of the 22 tables indicates that in 64 out of the 66 pairwise profile comparisons, a higher profile is more likely to choose a strategy in accordance with backward induction compared to a lower profile. The only two exceptions happen in games 2.2.2.2.2.2.2 and 4.2.2.2.2 where $P_{savvy-fast} < P_{savvy-slow}$.

Moreover, in 58 out of the 64 results supporting the profiling method, the difference between the higher and lower profiles is statistically significant at the 5% level or less, while in 51 of those 58 the difference is statistically significant at the 1% level or less. Therefore, the two-dimensional profile provides reliable predictive power.

Since scalar profiles are commonly used in the literature, it is only natural to study the predictive power of one-dimensional profiles: subjects divided into (unconditional) savvy/naive and, separately, (unconditional) fast/slow. In all 22 games, the savvy profile is more likely to choose a strategy in accordance with backward induction than the naive profile ($P_{savvy} > P_{naive}$). The difference between these two probabilities is always statistically significant at the 1% level or less. On the other hand, the fast profile is more likely to choose a strategy in accordance with backward induction than the slow profile in 15 out of 22 cases while the opposite happens in the other 7. Therefore, profiling unconditionally by total time (TT) is not reliable in terms of predictive power. Table XXVI presents these results.

[Table XXVI about here.]

To further study the predictive power of the proposed profiling, the following simple logit model is estimated for each of the 22 games:

$$Logit(Y) = \alpha + \beta X \tag{4}$$

Y is the dependent variable in the regression and captures whether the subject backward inducted ($Y_i = 1$) or did not backward induct ($Y_i = 0$), α is the intercept, and X is the $N \times 3$ matrix where N is the number of observations in each game and the three independent variables are $RRT1$, TT, and Seq. The independent variable Seq corresponds to the order in which a game appeared in the subject's sequence of games.

Table XXVII below shows the value of the estimated parameter for each of the three control variables (robust standard errors are in parenthesis), the pseudo R^2 of the regression, and the predicted power of the model in the last column. The predicted power of the model is calculated using the fitted value of the probabilities predicted by the model $\hat{p}_i = \Lambda(x_i\hat{\beta})$. Assuming a symmetric loss function, we assign $\hat{Y} = 1$ if the predicted value is $\hat{p}_i \geq 0.5$ and $\hat{Y} = 0$ if $\hat{p}_i < 0.5$.

[Table XXVII about here.]

First, the estimated coefficient of variables $RRT1$ and TT always have the expected signs: positive for $RRT1$ and negative for TT. All coefficients are significant at the 1% level or less, except two cases for TT where the coefficients are significant at the 5% level or less (games 3.2.3 and 4.2.2). The variable Seq is always positive and is statistically significant most of the time, indicating that the further in the sequence that a specific game appears, the higher the likelihood that the subject will behave according to backward induction.

Finally, the last column in Table XXVII shows the predictive power of the very simple model proposed to test the two-dimensional profile. The model correctly predicts whether a player behaved in accordance with backward induction or not between 73.61% and 94.75% percent of the time. The average percentage of correct predictions across games is 88.85%. These results confirm the good predictive power of the proposed profiling method.

To finalize the quantitative analysis and study the predictive impact of unconditional variables $RRT1$ and TT (i.e., scalar profiles as discussed above), two more logit models are studied. Compared to Model 1, in Model 2 variable TT is removed while in Model 3, it is $RRT1$ that is excluded. Results are presented in Table XXVIII below.

[Table XXVIII about here.]

Table XXVIII shows that in Model 2, $RRT1$ has the expected sign in every game. In fact, the estimated coefficients for $RRT1$ do not vary much with respect to those estimated in Model 1 where $RRT1$ and TT are used together as regressors. On the other hand, when TT is used alone as regressor in Model 3, it presents negative coefficients in only 15 out of 22 the games. This confirms that TT is helpful for profiling only when controlling by $RRT1$.

IV Conclusions

Blues and Reds, a mobile app for Android and iOS, was developed with the intention to take advantage of the omnipresence of mobile technology. The objective of *Blues and Reds* is to conduct mobile experiments; that is, experiments in which people install the app on their mobile devices and become the subjects of an experiment by using the app.

This article develops a novel method of profiling subjects playing dynamic games. A profile consists of two dimensions. The first dimension is about how people think and, quantified as the relative response time at the first round ($RRT1$), divides subjects into savvy (high $RRT1$) and naive (low $RRT1$). The second dimension is about how much people think and, measured by the total time spent on solving the game (TT), divides subjects into slow (high TT) and fast (low TT).

The predictive power of the proposed profiling is verified using a database generated by *Blues and Reds*. The database consists of 35,826 observations from 6,463 subjects located in 141 countries. The same empirical analysis is replicated in 22 various games. Each game is treated as a separate experiment and consists of the subject playing against Artificial

Intelligence in a dynamic game with perfect and complete information. The subject either wins or loses; winning is indicative of the subject backward inducting. For each subject and each game, *Blues and Reds* records response time at each round of the game and whether the subject wins.

Analyses show that the proposed profiling is quite accurate at predicting whether subjects backward induct. The probability of a subject backward inducting decreases in the following fashion: from savvy-fast to savvy-slow to naive-fast to naive-slow.

References

Agranov, M., A. Caplin, and C. Tergiman (2015): "Naive play and the process of choice in guessing games," *Journal of the Economic Science Association*, 1, 146–157.

Agranov, M., E. Potamites, A. Schotter, and C. Tergiman (2012): "Beliefs and endogenous cognitive levels: An experimental study," *Games and Economic Behavior*, 75, 449–463.

Alaoui, L. and A. Penta (2016): "Endogenous Depth of Reasoning," *Review of Economic Studies*, 83, 1297–1333.

Allred, S., S. Duffy, and J. Smith (2016): "Cognitive load and strategic sophistication," *Journal of Behavioral and Experimental Economics*, 51, 47–56.

Arad, A. and A. Rubinstein (2012): "The 11-20 Money Request Game: A Level-k Reasoning Study," *American Economic Review*, 102, 3561–3573.

Batzilis, D., S. Jaffe, S. Levitt, and J. A. List (2017): "How Facebook Can Deepen our Understanding of Behavior in Strategic Settings: Evidence from a Million Rock-Paper-Scissors Games," *working paper*.

Bayer, R. C. and L. Renou (2016): "Logical omniscience at the laboratory," *Journal of Behavioral and Experimental Economics*, 64, 41–49.

Benito-Ostolaza, J. M., P. Hernández, and J. A. Sanchis-Llopis (2016): "Do individuals with higher cognitive ability play more strategically?" *Journal of Behavioral and Experimental Economics*, 64, 5–11.

Bosch-Domènech, A., J. G. Montalvo, R. Nagel, and A. Satorra (2002): "One, Two, (Three), Infinity, ... : Newspaper and Lab Beauty-Contest Experiments," *American Economic Review*, 92, 1687–1701.

Brañas-Garza, P., T. García-Muñoz, and R. H. González (2012): "Cognitive effort in the Beauty Contest Game," *Journal of Economic Behavior & Organization*, 83, 254–260.

Burchardi, K. B. and S. P. Penczynski (2014): "Out of Your Mind: Eliciting Individual Reasoning in One Shot Games," *Games and Economic Behavior*, 84, 39–57.

Burks, S. V., J. P. Carpenter, L. Goette, and A. Rustichini (2009): "Cognitive Skills Affect Economic Preferences, Strategic Behavior, and Job Attachment," *Proceedings of the National Academy of Sciences*, 106, 7745–7750.

Burnham, T. C., D. Cesarini, M. Johannesson, P. Lichtenstein, and B. Wallace (2009): "Higher Cognitive Ability is Associated with Lower Entries in a p-beauty Contest," *Journal of Economic Behavior & Organization*, 72, 171–175.

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2004): "A Cognitive Hierarchy Model of Games," *Quarterly Journal of Economics*, 119, 861–898.

Carpenter, J., M. Graham, and J. Wolf (2013): "Cognitive Ability and Strategic Sophistication," *Games and Economic Behavior*, 80, 115–130.

Chen, Y., G. Y. Jeon, and Y.-M. Kim (2014): "A day without a search engine: an experimental study of online and offline searches," *Experimental Economics*, 17, 512–536.

Chen, Y. and J. Konstan (2015): "Online field experiments: a selective survey of methods," *Journal of the Economic Science Association*, 1, 29–42.

Clithero, J. A. (2016): "Response Times in Economics: Looking Through the Lens of Sequential Sampling Models," *working paper*.

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

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

Costa-Gomes, M. A. and G. Weizsäcker (2008): "Stated Beliefs and Play in Normal-Form Games," *Review of Economic Studies*, 75, 729–762.

Duffy, S. and J. Smith (2014): "Cognitive Load in the Multi-player Prisoner's Dilemma Game: Are There Brains in Games?" *Journal of Behavioral and Experimental Economics*, 51, 47–56.

Evans, A. M., K. D. Dillon, and D. G. Rand (2015): "Fast But Not Intuitive, Slow But Not Reflective: Decision Conflict Drives Reaction Times in Social Dilemmas," *Journal of Experimental Psychology: General*, 144, 951–966.

Fehr, D. and S. Huck (2016): "Who Knows It is a Game? On Strategic Awareness and Cognitive Ability," *Experimental Economics*, 19, 713–726.

Georganas, S., P. J. Healy, and R. A. Weber (2015): "On the Persistence of Strategic Sophistication," *Economic Theory*, 159, 369–400.

Gill, D. and V. Prowse (2016): "Cognitive Ability, Character Skills, and Learning to Play Equilibrium: A Level-*k* Analysis," *Journal of Political Economy*, 124, 1619–1676.

——— (2017): "Strategic Complexity and the Value of Thinking," *working paper*.

Hanaki, N., N. Jacquemet, S. Luchini, and A. Zylbersztejn (2016): "Cognitive Ability and the Effect of Strategic Uncertainty," *Theory and Decision*, 81, 101–121.

Hargreaves Heap, S., D. Rojo Arjona, and R. Sugden (2014): "How Portable Is Level-0 Behavior? A Test of Level-*k* Theory in Games With Non-Neutral Frames," *Econometrica*, 82, 1133–1151.

Ho, T.-H., C. Camerer, and K. Weigelt (1998): "Iterated Dominance and Iterated Best Response in Experimental "p-Beauty Contests"," *American Economic Review*, 88, 947–969.

Ho, T.-H. and X. Su (2013): "A Dynamic Level-*k* Model in Sequential Games," *Management Science*, 59, 452–469.

Kahneman, D. (2013): *Thinking, Fast and Slow*, Farrar, Straus and Giroux.

Kiss, H., I. Rodriguez-Larac, and A. Rosa-García (2016): "Think Twice Before Running! Bank Runs and Cognitive Abilities," *Journal of Behavioral and Experimental Economics*, 64, 12–19.

Liu, T. X., J. Yang, L. A. Adamic, and Y. Chen (2014): "Crowdsourcing with All-Pay Auctions: A Field Experiment on Taskcn," *Management Science*, 60, 2020–2037.

Lohse, J., T. Goeschl, and J. H. Diederich (2017): "Giving is a Question of Time: Response Times and Contributions to an Environmental Public Good," *Environmental and Resource Economics*, 67, 455–477.

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

Palacios-Huerta, I. and O. Volij (2009): "Field Centipedes," *American Economic Review*, 9, 1619–1635.

Penczynski, S. P. (2016): "Strategic Shinking: The Influence of the Game," *Journal of Economic Behavior & Organization*, 128, 72–84.

Piovesan, M. and E. Wengstr'om (2009): "Fast or fair? A study of response times," *Economics Letters*, 105, 193–196.

Rand, D. G., J. D. Greene, and M. A. Nowak (2012): "Spontaneous Giving and Calculated Greed," *Nature*, 489, 427–430.

Rubinstein, A. (2006): "Dilemmas of an Economic Theorist," *Econometrica*, 74, 865–883.

——— (2007): "Instinctive and Cognitive Reasoning: A Study of Response Times," *Economic Journal*, 117, 1243–1259.

——— (2013): "Response Time and Decision Making: An Experimental Study," *Judgment and Decision Making*, 8, 540–551.

——— (2016): "A Typology of Players: Between Instinctive and Contemplative," *Quarterly Journal of Economics*, 131, 859–890.

Rydval, O., A. Ortmann, and M. Ostatnicky (2009): "Three Very Simple Games and What It Takes to Solve Them," *Journal of Economic Behavior & Organization*, 72, 589–601.

Schotter, A. and I. Trevino (2014): "Is Response Time Predictive of Choice? An Experimental Study of Threshold Strategies," *WZB Discussion Paper SP II 2014-305.*

Shapiro, D., X. Shi, and A. Zillante (2014): "Level-k Reasoning in a Generalized Beauty Contest," *Games and Economic Behavior*, 86, 308–329.

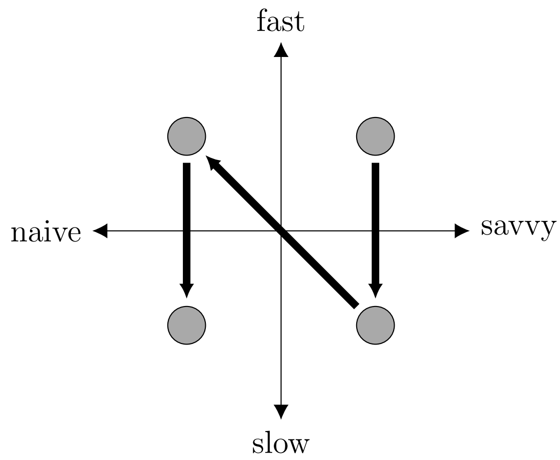
Spiliopoulos, L. and A. Ortmann (2017): "The BCD of Response Time Analysis in Experimental Economics," *Experimental Economics*, 47, 1–55.

Stahl, D. O. and P. W. Wilson (1994): "Experimental Evidence on Players' Models of Other Players," *Journal of Economic Behavior & Organization*, 25, 309–327.

——— (1995): "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, 10, 218–254.

Wang, J., M. Spezio, and C. F. Camerer (2010): "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games," *American Economic Review*, 100, 984–1007.

Figure I: Two-dimensional profiling of players in dynamic games.



Notes. Each circle represents a two-dimensional profile. Probability of subject not making a mistake (i.e., behaving consistently with backward induction) decreases with the direction of arrows.

Figure II: An example of game in *Blues and Reds*.

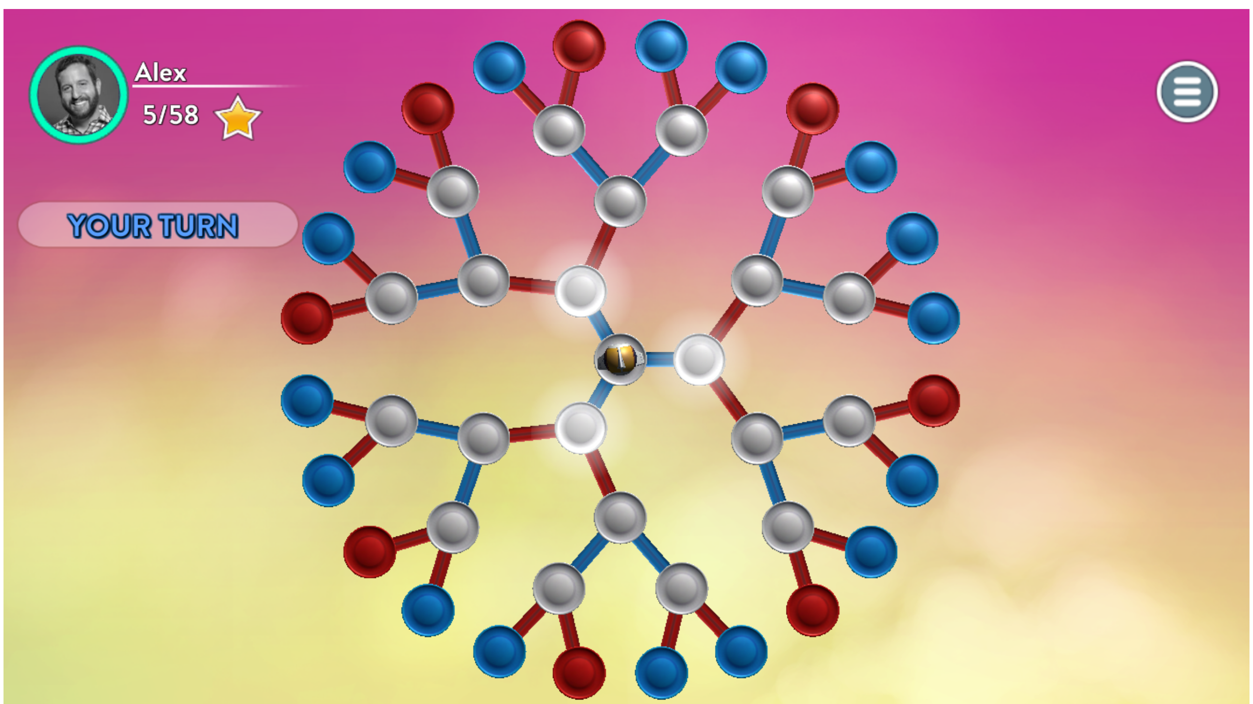
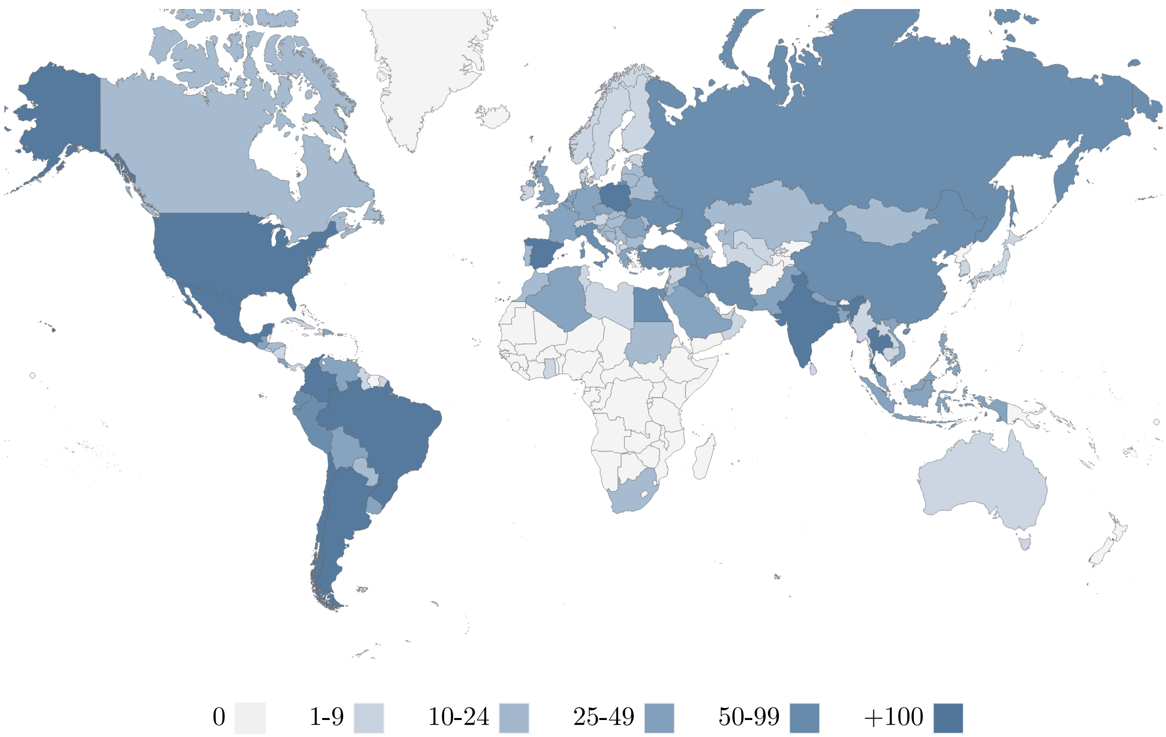


Figure III: Geographical distribution of subjects.



Notes. For 5,746 out of 6,463 subjects, it was possible to identify subject's location by her device's IP. This figure is created for the base 5,746 subjects. Each color depicts the number of subjects from a given country. Heat map created with `https://public.tableau.com`.

Table I: Summary statistics.

game	N	$\%Win$	$RRT1$				TT			
			Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max
2.2.2	1,638	94%	0.71	0.11	0.13	0.93	19.05	7.17	5	50
2.2.3	1,729	94%	0.74	0.10	0.06	0.93	21.07	7.22	6	49
2.3.2	1,630	92%	0.71	0.11	0.10	0.92	21.04	7.86	8	50
2.3.3	1,637	93%	0.78	0.11	0.13	0.96	22.98	9.82	8	62
3.2.2	1,666	91%	0.75	0.11	0.15	0.94	21.03	7.76	7	49
3.3.2	1,647	90%	0.77	0.12	0.12	0.96	23.18	9.56	8	61
3.2.3	1,628	91%	0.77	0.11	0.10	0.96	21.94	8.44	8	56
3.3.3	1,638	90%	0.80	0.12	0.18	0.96	25.64	9.63	5	58
4.2.2	1,717	89%	0.77	0.12	0.11	0.95	22.68	9.79	6	62
2.2.2.2	1,660	67%	0.72	0.19	0.12	0.96	30.10	14.60	6	85
2.2.2.3	1,610	79%	0.78	0.15	0.19	0.98	33.24	17.01	6	93
2.2.3.2	1,674	77%	0.77	0.16	0.12	0.98	34.55	16.91	9	90
2.3.2.2	1,606	56%	0.73	0.20	0.12	0.98	35.98	18.44	6	101
3.2.2.2	1,575	70%	0.79	0.16	0.14	0.98	38.44	20.65	5	118
2.2.2.4	1,602	83%	0.79	0.13	0.18	0.96	32.57	15.96	7	87
2.2.4.2	1,673	73%	0.77	0.17	0.07	0.97	40.44	20.88	7	112
2.4.2.2	1,641	81%	0.78	0.14	0.18	0.96	35.84	17.94	9	96
4.2.2.2	1,614	70%	0.79	0.16	0.14	0.98	38.63	21.69	8	121
2.2.2.2.2	1,545	72%	0.69	0.19	0.06	0.95	59.43	32.15	10	184
3.2.2.2.2	1,550	67%	0.68	0.22	0.04	0.97	66.26	42.55	12	235
4.2.2.2.2	1,566	48%	0.67	0.23	0.02	0.98	95.08	78.10	10	534
2.2.2.2.2.2	1,580	47%	0.61	0.21	0.08	0.97	81.55	60.14	11	328

Notes. N denotes the number of subjects who played a given game. $\%Win$ is the percentage of subjects who won (i.e., behaved consistently with backward induction). For $RRT1$ and TT, this table provides the mean, standard deviation, and minimal and maximal values.

Table II: Response times of four fictional subjects in the game 3.2.2.2.

subject	$RT1$	$RT3$	TT
Ann	15	5	20
Bob	30	10	40
Chris	8	12	20
David	16	24	40

Table III: Example of table with data.

	naive	savvy
fast	$P_{naive-fast}$	$P_{savvy-fast}$
slow	$P_{naive-slow}$	$P_{savvy-slow}$

(1) t-stat, (2) t-stat, (3) t-stat

Notes. (1), (2), and (3) provide t-statistics for the three inequalities: (1) $P_{savvy-fast} > P_{savvy-slow}$, (2) $P_{savvy-slow} > P_{naive-fast}$, and (3) $P_{naive-fast} > P_{naive-slow}$. Stars indicate one-tail test level of significance (***1%, **5%, *10%).

Table IV: Game 2.2.2

	naive	savvy
fast	95.88%	99.50%
slow	80.73%	97.83%

(1) 2.10**, (2) 1.64*, (3) 6.80***

Table V: Game 2.2.3

	naive	savvy
fast	94.57%	100.00%
slow	81.88%	98.17%

(1) 2.85***, (2) 2.86***, (3) 5.93***

Table VI: Game 2.3.2

	naive	savvy
fast	94.76%	99.07%
slow	75.80%	96.79%

(1) 2.29**, (2) 1.45*, (3) 7.69***

Table VII: Game 2.3.3

	naive	savvy
fast	96.50%	99.75%
slow	78.86%	98.53%

(1) 1.90**, (2) 1.85**, (3) 8.04***

Table VIII: Game 3.2.2

	naive	savvy
fast	93.42%	99.31%
slow	69.81%	97.86%

(1) 1.78**, (2) 3.22***, (3) 8.87***

Table IX: Game 3.3.2

	naive	savvy
fast	90.34%	99.48%
slow	75.29%	95.67%

(1) 3.49***, (2) 3.07***, (3) 5.97***

Table X: Game 3.2.3

	naive	savvy
fast	91.95%	99.26%
slow	75.77%	98.21%

(1) 1.33*, (2) 4.16***, (3) 6.51***

Table XI: Game 3.3.3

	naive	savvy
fast	90.14%	99.76%
slow	71.14%	97.26%

(1) 2.94***, (2) 4.24***, (3) 7.05***

Table XII: Game 4.2.2

	naive	savvy
fast	92.35%	99.05%
slow	68.79%	97.85%

(1) 1.45*, (2) 3.66***, (3) 9.09***

Table XIII: Game 2.2.2.2

	naive	savvy
fast	63.22%	97.12%
slow	13.13%	93.89%

(1) 2.24**, (2) 11.58***, (3) 17.36***

Table XIV: Game 2.2.2.3

	naive	savvy
fast	75.12%	99.49%
slow	46.43%	94.85%

(1) 4.02***, (2) 8.28***, (3) 8.71***

Table XV: Game 2.2.3.2

	naive	savvy
fast	70.49%	99.29%
slow	43.12%	96.61%

(1) 2.73***, (2) 10.77***, (3) 8.32***

Table XVI: Game 2.3.2.2

	naive	savvy
fast	31.07%	96.53%
slow	8.52%	90.82%

(1) 3.31***, (2) 22.05***, (3) 8.42***

Table XVII: Game 3.2.2.2

	naive	savvy
fast	54.86%	98.97%
slow	31.35%	93.22%

(1) 4.23***, (2) 13.75***, (3) 6.85***

Table XVIII: Game 2.2.2.4

	naive	savvy
fast	75.25%	98.04%
slow	61.38%	96.71%

(1) 1.18, (2) 9.25***, (3) 4.25***

Table XIX: Game 2.2.4.2

	naive	savvy
fast	68.60%	98.80%
slow	28.98%	94.76%

(1) 3.34***, (2) 10.34***, (3) 12.46***

Table XX: Game 2.4.2.2

	naive	savvy
fast	74.24%	97.77%
slow	52.69%	96.66%

(1) 0.97, (2) 9.77***, (3) 6.53***

Table XXI: Game 4.2.2.2

	naive	savvy
fast	60.10%	96.99%
slow	31.00%	93.15%

(1) 2.53***, (2) 12.08***, (3) 8.66***

Table XXII: Game 2.2.2.2.2

	naive	savvy
fast	70.44%	97.69%
slow	31.07%	89.82%

(1) 4.57***, (2) 6.96***, (3) 11.88***

Table XXIII: Game 3.2.2.2.2

	naive	savvy
fast	53.26%	94.86%
slow	29.08%	91.19%

(1) 2.01^{**}, (2) 12.93^{***}, (3) 7.04^{***}

Table XXIV: Game 4.2.2.2.2

	naive	savvy
fast	28.83%	76.28%
slow	7.42%	79.28%

(1) -1.01, (2) 16.41^{***}, (3) 8.09^{***}

Table XXV: Game 2.2.2.2.2.2

	naive	savvy
fast	29.59%	61.36%
slow	21.66%	74.94%

(1) -4.14, (2) 14.27^{***}, (3) 2.56^{***}

Table XXVI: One-dimensional profiles.

game	savvy	naive	fast	slow	game	savvy	naive	fast	slow
2.2.2	98.65%	88.79%	97.34%	89.64%	2.2.3.2	97.96%	56.50%	78.13%	76.25%
	(8.40***)		(6.28***)			(23.28***)		(0.92)	
2.2.3	99.06%	88.27%	95.83%	91.21%	2.3.2.2	93.71%	19.98%	58.60%	54.23%
	(9.50***)		(3.90***)			(44.74***)		(1.76*)	
2.3.2	97.96%	85.80%	96.47%	88.09%	3.2.2.2	96.07%	43.33%	67.44%	71.95%
	(9.13***)		(6.50***)			(27.78***)		(−1.95*)	
2.3.3	99.14%	87.45%	97.06%	89.19%	2.2.2.4	97.38%	68.46%	82.81%	83.11%
	(9.73***)		(6.30***)			(16.63***)		(−0.16)	
3.2.2	98.59%	82.64%	94.27%	87.03%	2.2.4.2	96.78%	48.62%	75.76%	69.81%
	(11.48***)		(5.06***)			(26.24***)		(2.74***)	
3.3.2	97.55%	82.99%	93.12%	86.67%	2.4.2.2	97.21%	63.94%	78.38%	82.76%
	(10.48***)		(4.37***)			(18.74***)		(−2.24**)	
3.2.3	98.75%	83.75%	93.95%	88.26%	4.2.2.2	95.05%	45.66%	67.54%	73.16%
	(11.19***)		(4.05***)			(25.80***)		(−2.47**)	
3.3.3	98.54%	80.81%	93.47%	85.82%	2.2.2.2.2	93.79%	50.91%	73.87%	70.85%
	(12.31***)		(5.11***)			(21.45***)		(1.32)	
4.2.2	98.42%	79.90%	94.11%	85.25%	3.2.2.2.2	93.03%	41.03%	59.69%	74.36%
	(12.75***)		(6.15***)			(26.12***)		(−6.21***)	
2.2.2.2	95.52%	38.08%	73.92%	58.88%	4.2.2.2.2	77.78%	18.14%	35.48%	60.28%
	(31.39***)		(6.55***)			(29.42***)		(−10.14***)	
2.2.2.3	97.13%	61.23%	81.43%	76.66%	2.2.2.2.2.2	68.14%	25.60%	35.71%	58.31%
	(19.81***)		(2.35**)			(18.72***)		(−9.23***)	

Notes. Subjects with $RRT1$ below and above the median $RRT1$ are called savvy and naive, respectively. Subjects with TT below and above the median are called fast and slow, respectively. P_i is the percentage of subjects of profile i who behaved in line with backward induction. T-statistics for the differences $P_{savvy} - P_{naive}$ and $P_{fast} - P_{slow}$ are in the brackets. Stars indicates two-tail test the significance (***1%, **5%, *10%).

Table XXVII: Predictive power of the profiling method using logit regressions.

game	N	$RRT1$	TT	Seq	pseudo R^2	correctly predicted
2.2.2	1,638	10.06	-0.07	0.14	0.34	94.75%
		(0.98)	(0.02)	(0.03)		
2.2.3	1,729	12.28	-0.06	0.12	0.36	94.33%
		(1.13)	(0.02)	(0.03)		
2.3.2	1,630	10.21	-0.07	0.06	0.33	93.68%
		(0.98)	(0.01)	(0.02)		
2.3.3	1,637	12.59	-0.06	0.12	0.44	94.75%
		(1.36)	(0.02)	(0.04)		
3.2.2	1,666	12.54	-0.05	0.09	0.39	93.10%
		(1.11)	(0.02)	(0.02)		
3.3.2	1,647	10.37	-0.05	0.08	0.32	91.20%
		(0.97)	(0.01)	(0.02)		
3.2.3	1,628	11.93	-0.04	0.06	0.36	92.57%
		(0.99)	(0.02)	(0.02)		
3.3.3	1,638	13.29	-0.04	0.08	0.43	93.41%
		(1.18)	(0.01)	(0.03)		
4.2.2	1,717	13.08	-0.03	0.11	0.44	93.01%
		(0.93)	(0.01)	(0.02)		
2.2.2.2	1,660	15.34	-0.05	0.04	0.54	89.04%
		(1.04)	(0.01)	(0.01)		
2.2.2.3	1,610	13.43	-0.04	0.08	0.44	88.07%
		(0.98)	(0.01)	(0.02)		
2.2.3.2	1,674	13.89	-0.03	0.05	0.47	88.77%
		(1.01)	(0.01)	(0.02)		
2.3.2.2	1,606	19.62	-0.04	0.03	0.61	90.04%
		(1.44)	(0.01)	(0.01)		
3.2.2.2	1,575	18.78	-0.04	0.05	0.53	89.21%
		(1.40)	(0.01)	(0.01)		
2.2.2.4	1,602	11.62	-0.02	0.08	0.34	87.45%
		(0.92)	(0.01)	(0.02)		
2.2.4.2	1,673	14.91	-0.04	0.02	0.50	87.87%
		(1.01)	(0.00)	(0.01)		
2.4.2.2	1,641	13.24	-0.03	0.05	0.38	87.63%
		(1.02)	(0.01)	(0.02)		
4.2.2.2	1,614	19.24	-0.04	0.00	0.50	87.24%
		(1.32)	(0.01)	(0.01)		
2.2.2.2.2	1,545	10.59	-0.03	0.02	0.36	82.39%
		(0.63)	(0.00)	(0.01)		
3.2.2.2.2	1,550	9.74	-0.02	0.03	0.38	82.58%
		(0.61)	(0.00)	(0.01)		
4.2.2.2.2	1,566	9.95	-0.01	0.00	0.37	80.08%
		(0.54)	(0.00)	(0.01)		
2.2.2.2.2.2	1,580	6.54	0.00	0.02	0.23	73.61%
		(0.40)	(0.00)	(0.01)		

Notes. The table shows the results from the estimation of Model 1 for each game. Robust standard errors are in parenthesis.

Table XXVIII: Logit regressions for RRT1 and TT separately.

game	N	Model 2			Model 3		
		$RRT1$	Seq	pseudo R^2	TT	Seq	pseudo R^2
2.2.2	1,638	11.13	0.14	0.32	-0.09	0.13	0.13
		(0.92)	(0.03)		(0.01)	(0.03)	
2.2.3	1,729	12.78	0.11	0.34	-0.07	0.06	0.06
		(1.07)	(0.03)		(0.01)	(0.02)	
2.3.2	1,630	11.28	0.06	0.30	-0.08	0.06	0.11
		(0.95)	(0.02)		(0.01)	(0.02)	
2.3.3	1,637	12.86	0.11	0.41	-0.06	0.12	0.11
		(1.19)	(0.03)		(0.01)	(0.03)	
3.2.2	1,666	13.24	0.09	0.37	-0.06	0.08	0.07
		(1.12)	(0.02)		(0.01)	(0.02)	
3.3.2	1,647	10.73	0.08	0.30	-0.05	0.08	0.07
		(0.92)	(0.02)		(0.01)	(0.02)	
3.2.3	1,628	12.52	0.06	0.35	-0.05	0.05	0.05
		(0.96)	(0.02)		(0.01)	(0.02)	
3.3.3	1,638	13.19	0.09	0.42	-0.04	0.09	0.06
		(1.07)	(0.02)		(0.01)	(0.02)	
4.2.2	1,717	13.48	0.11	0.43	-0.05	0.09	0.09
		(0.92)	(0.02)		(0.01)	(0.02)	
2.2.2.2	1,660	14.05	0.05	0.51	-0.02	0.06	0.05
		(0.78)	(0.01)		(0.00)	(0.01)	
2.2.2.3	1,610	11.76	0.08	0.41	-0.01	0.09	0.05
		(0.73)	(0.02)		(0.00)	(0.01)	
2.2.3.2	1,674	13.05	0.05	0.45	0.00	0.09	0.04
		(0.85)	(0.02)		(0.00)	(0.01)	
2.3.2.2	1,606	16.69	0.04	0.58	0.00	0.07	0.04
		(1.03)	(0.01)		(0.00)	(0.01)	
3.2.2.2	1,575	15.59	0.05	0.50	0.01	0.08	0.04
		(0.94)	(0.01)		(0.00)	(0.01)	
2.2.2.4	1,602	11.00	0.08	0.33	0.01	0.09	0.05
		(0.75)	(0.02)		(0.00)	(0.01)	
2.2.4.2	1,673	12.81	0.02	0.44	-0.01	0.06	0.03
		(0.71)	(0.01)		(0.00)	(0.01)	
2.4.2.2	1,641	11.42	0.05	0.36	0.01	0.08	0.04
		(0.69)	(0.01)		(0.00)	(0.01)	
4.2.2.2	1,614	15.60	0.01	0.46	0.00	0.05	0.02
		(0.85)	(0.01)		(0.00)	(0.01)	
2.2.2.2.2	1,545	7.72	0.02	0.28	0.00	0.05	0.02
		(0.43)	(0.01)		(0.00)	(0.01)	
3.2.2.2.2	1,550	7.78	0.03	0.35	0.01	0.07	0.05
		(0.39)	(0.01)		(0.00)	(0.01)	
4.2.2.2.2	1,566	8.07	0.00	0.35	0.01	0.04	0.05
		(0.42)	(0.01)		(0.00)	(0.01)	
2.2.2.2.2.2	1,580	5.78	0.02	0.22	0.01	0.05	0.06
		(0.32)	(0.01)		(0.00)	(0.01)	

Notes. The table shows the results from the estimation of Model 2 and Model 3 for each game. Robust standard errors are in parenthesis.