Measuring tree complexity with response times*

Konrad Grabiszewski† Alex Horenstein ‡

Abstract

Game-theoretic trees vary in complexity. This paper introduces the concept of graph-based complexity and relies on the subjects' behavior to empirically derive a measure of tree complexity. Data comes from the mobile app *Blues and Reds*, designed specifically to conduct experiments. The sample consists of 6,637 subjects from 143 countries who play 27 different dynamic games. Based on subjects' response times, we find that two measures – the average response time spent at the first round and the average total time spent solving the tree – are the best candidates for the empirical measure of tree complexity. We focus on two-person, finite, zero-sum dynamic games with perfect and complete information.

JEL Codes: C72, C73, C80, C90.

Key words: tree, complexity, backward induction, 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 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 Economic Science Association, 2018 Spring Midwest Economic Theory, 2018 D-TEA, 2018 FUR, 2018 WIEM) and seminars (ITAM, Keio University, Kindai University, Kansai University, KAUST, University of Miami, MBSC).

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

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

1 Introduction

The standard way of depicting a dynamic problem is with a tree; we rely on backward induction to solve the tree. While every tree is equally complex from the algorithmic perspective, it is not the case from a human standpoint: some trees are more complex than others. Measuring tree complexity remains an open challenge in the economics literature, which we address in this paper.

Among various factors affecting tree complexity, our focus is on the graph itself. To that end, we begin with a notion of the *graph-based complexity* according to which a tree X is less complex than tree Y if Y can be constructed by adding branches and nodes to X. However, the graph-based measure of complexity is incomplete in a sense that it is unable to rank all pairs of trees. To solve this issue, we propose the concept of the empirical measure of tree complexity which satisfies the properties described below.

1. Complete order: we want this measure to be able to rank any pair of trees in terms of their complexity.

2. The best extension of the graph-based measure: since many measures might be complete, we focus on a measure that agrees with the graph-based measure for more pairwise comparisons than any other measure.

3. Empirical: refers to the measure which relies on data generated by the subjects. Their behavior indicates the complexity ranking of trees.

This paper offers three contributions. First, the proposed measure is of practical value for empirical studies in which it is necessary to control for the complexity of dynamic games the subjects play. For instance, this would be the case of studies – like ours – in which behavior is observed in various settings and it is suspected that complexity affects the said behavior. To date, a control variable for tree complexity has not been proposed; our paper fills the void. As an example, we show that not controlling for complexity might lead to the omitted-variable bias problem.

Second, our study of tree complexity sheds new light on the large literature devoted to analyzing backward induction. The generally accepted conclusion is that people behave inconsistently with what backward induction predicts. Yet, the relevant and still debated question is why they do so. Subjects' imperfect strategic reasoning is the main explanation found in the literature (e.g., Burks et al. (2009), Palacios-Huerta and Volij (2009), Carpenter

et al. (2013), Benito-Ostolaza et al. (2016), and Kiss et al. (2016)). Tree complexity is a novel reason for subjects to violate backward induction; namely, some games are just too complex for some subjects to solve.

Third, we analyze how a tree structure affects its complexity. We find that complexity increases with tree length and width. However, the length has a relatively larger impact on complexity.

To collect the data, we conducted a mobile experiment, that is an experiment which takes place on subjects' smartphones and tablets. We hired a team of professional app developers who created *Blues and Reds*, a free mobile app available globally in English, Spanish, Chinese simplified, and Chinese traditional for iOS and Android devices. On the website dedicated to *Blues and Reds* (`http://www.bluesandreds.com/`), the readers can find an additional description of the app as well as the links to Google Play and the App Store for installing.[1]

This innovative approach to data collection is inspired by the worldwide proliferation of mobile technology, smartphones, and tablets. Nowadays, it is impossible to imagine an average citizen of the world going through their daily routines without a mobile device. With more people owning more devices and using them more frequently, the mobile revolution offers attractive opportunities to conduct global experiments.[2] This paper explores these opportunities for academic research.

In this paper, we use data generated by 27 dynamic games from *Blues and Reds*. Each game is a two-person, zero-sum, no-tie, finite dynamic game with perfect and complete information. In terms of graphical visualization, games in *Blues and Reds* resemble classical game-theoretic trees. Games vary in the number of rounds (from 2 to 6) and the number of actions at each round.

The selection of the specific games in the app was driven by the goal to have as many various trees as possible. This allowed us to build a large data set. Technical limitations – more precisely, the size of an average smartphone – did not permit a clear visualization on the screen of a whole tree with more than six rounds or more branches per round.

[1] In March 2018, immediately after the bulk of our data was collected, *Blues and Reds* had 4.1 (out of 5) stars in the ranking on Google Play, placing it among very popular titles produced by multi-billion-dollar companies: *Candy Crush Saga* (4.4 stars, Activision Blizzard), *Tetris* (4.0 stars, Electronic Arts), and *Super Mario Run* (3.8 stars, Nintendo).

[2] The 2020 COVID-19 pandemic put the whole world on lockdown, including experimental labs located on university campuses. Traditional experiments had to be placed on hold. Writing this paper during the pandemic, we expect that, in the post-pandemic world, some restrictions will not be lifted (especially those related to maintaining social distance) making it difficult if not impossible to run traditional lab sessions. These restrictions have no impact on conducting mobile experiments which require no physical lab, and the contact between subjects and the experimenters is purely virtual.

In each game, a human subject plays against an algorithm that we programmed to backward induct. The subject's goal is to win, an outcome possible in each game as long as the subject follows a path of actions as prescribed by backward induction. Deviating from that path guarantees a subject's loss as the algorithm exploits the subject's mistakes.

The database used in this paper consists of 6,637 subjects from 143 countries. For each game and each subject, *Blues and Reds* records the following variables: (i) whether or not a subject wins, and (ii) her response time (RT), measured in seconds, at each round.

As it is standard in the literature, subject's RT is considered as a proxy of their cognitive effort; longer RT implies higher effort (e.g., Ofek et al. (2007), Rubinstein (2007), Fehr and Rangel (2011), Rubinstein (2013), Krajbich et al. (2014), Caplin and Martin (2016), Clithero (2018), Rubinstein (2016), Gill and Prowse (2017), Spiliopoulos and Ortmann (2017), Recalde et al. (2018), and Enke and Zimmermann (2019)). At the same time, subject's winning/losing is the outcome of their exerting that cognitive effort.

Using these variables, we construct three candidate variables for the empirical measure of tree complexity:

1. the average response time that subjects spend at the first round of a tree ($ART1$),
2. the average total time that subjects spend solving the tree (ATT), and
3. the percentage of subjects who did not backward induct ($\%NOT.BI$).

Each candidate satisfies two out of the three required properties listed above (complete ranking and based on behavioral data). Hence, to determine the best measure, we analyze how well the three candidates extend the graph-based measure.

With 27 different games, there are 351 pairwise tree comparisons. For 136 of these pairs, it is possible to determine which tree is more complex than another in terms of their graphs. Hence, we test how well the proposed candidates replicate the graph-based measure for these 136 pairs. We find that $ART1$ and ATT do an equally good job as they agree with the graph-based ranking for 133 and 132 pairs, respectively. At the same time, $\%NOT.BI$ is a far inferior measure as it agrees with the graph-based measure for only 105 pairs. Consequently, we select $ART1$ and ATT as the empirical measure of tree complexity.[3]

From the methodological perspective, this paper belongs to the line of research that relies on non-standard methods to gather the data, like newspapers-based experiments (e.g.,

[3]Gill and Prowse (2017) also measure the complexity of games and, like us, also rely on RTs; however, their study is in the context of static games.

Bosch-Domènech et al. (2002)) and online experiments (e.g., Ariel Rubinstein's `https://arielrubinstein.org/gt/`, Chen et al. (2014), and Liu et al. (2014), and Chen and Konstan (2015)).

The sole objective of *Blues and Reds* is to conduct economic experiments and collect data. Various projects are based on data from *Blues and Reds*. In particular, in Grabiszewski and Horenstein (2020a), we study the monotonicity of the relationship between skills and effort.

In a given game, to measure the subject's skills, we compute the relative response time at the first round; that is, $RRT1 = \frac{RT1}{TT}$, where $RT1$ is the subject's response time at the first round and TT is the total time the subject spent on solving a game. Given that the correctly implemented backward induction requires investing cognitive effort at the first round, we argue that higher $RRT1$ indicates higher backward-inducting skills. When it comes to effort, we follow the literature and measure effort as TT.

We find that effort is decreasing in skills when the level of skills is low enough but increasing for skills high enough. In other words, the relationship is U-shaped. In Grabiszewski and Horenstein (2020a), we focus on players' behavior in trees while this paper focuses on the trees themselves. From this perspective, this article and Grabiszewski and Horenstein (2020a) complement each other.

Grabiszewski and Horenstein (2020b) is another paper based on data from *Blues and Reds*. Here, we study the problem of perception. Before players backward induct and decide on their strategies, they need to understand the game they participate in. Correct game-form recognition is critical for a success and precedes solving a game. We find that, in general, people struggle with the crucial step of understanding games they play. Our data also indicates that, in terms of success rate, correct backward inducting is easier than correct game-form recognition, suggesting that people might only appear as if failing at backward induction while they actually correctly solve incorrectly constructed trees.

The rest of this paper is organized as follows. Section 2 discusses the notion of graph-based complexity. Section 3 describes the experiment. Section 4 contains the search for the empirical measure of complexity. Section 5 studies the connection between a tree structure and tree complexity. Section 6 shows that not controlling for complexity potentially leads to the omitted-variable bias problem. Section 7 concludes. Appendix A contains the screenshots of all the games from the experiment. Appendix B describes *Blues and Reds* as a mobile game and includes experimental instructions. Appendices C and E include additional empirical results. Appendix D includes additional data.

2 Graph-based complexity

In this paper, we think of complexity as reflecting the difficulty of accomplishing a specific task. For finite dynamic games with perfect and complete information, complexity captures how hard it is to solve a game by correctly applying backward induction.

The process of backward inducting begins with analyzing a tree at its end and moves up towards the initial node. At each stage, there is a comparison of available alternatives to find the best one. The more alternatives there are the more complex the comparison is because the process must conduct more operations.[4]

Looking at complexity from the perspective of implementing backward induction leads to the concept of graph-based complexity: a tree X is graph-based less complex than tree Y if Y can be constructed by adding branches and nodes to X. From this perspective, in Figure 1, Tree A is less complex than both Trees B and C. Note that in Figure 1, it is not a coincidence that the trees lack payoffs. This is because for the specific payoffs do not make a tree more or less complex for backward induction.

[Figure 1 about here.]

However, graph-based complexity measure does not generate the complete ranking of trees. For example, in Figure 1, it is not possible to use this measure to compare Trees B and C. This incompleteness is precisely why this paper advances the idea of the empirical measure of complexity; i.e., the best and complete extension of graph-based measure.

Naturally, graph-based is not the only approach to complexity; other factors also play a significant role. Who the opponents are – or rather their rationality – is another variable affecting complexity. We can think about the concept of opponent-based complexity: some players make a game less complex than others. For instance, playing chess with a rookie seems less complex than against a chess master; the former is easier to beat.

Yet, it might be that less rational opponents make a game more complex. To elaborate, consider Trees D and E in Figure 2. In each tree, if Ann believes that her opponent is rational, she should go right as this yields a certain payoff of 3. However, if she is allowed for a possibility of the opponent making a mistake, then going left might be a better option.

[4] In computer science, the problem of finding the best alternative from a finite set is a special case of what is called the selection problem, a problem that entered the realm of mathematics in the 1930s. Complexity of such a problem increases with the cardinality of the set a selection comes from (e.g., Cormen et al. (2009)).

[Figure 2 about here.]

We argue that in Figure 2 playing against an irrational opponent is more complex. Suppose that Ann knows that Bob is rational. She assigns probability $p = 0$ to Bob making a mistake. In this case, she just needs to backward induct.

However, she knows that Chris is not fully rational. With probability $p > 0$, Chris is going to make a mistake. First, she has to figure out how irrational Chris is. To that end, her cognitive investment increases as she estimates the value of p. If she obtains a point estimate of p, she follows the expected utility theory to choose between left and right; compared to the game against Bob, this is clearly a more complex deliberation. It can become even more complex if her analysis yields a range of probabilities instead of just one point. In the end, playing against rational Bob is simpler and easier than facing irrational Chris.

Trees D and E in Figure 2 are equally complex from the graph-based perspective. However, from Ann's perspective these two games need not be equally difficult to solve as who Bob and Chris are determines the opponents-based complexity.

Payoff distribution is also a potential source of complexity. In Figure 3, Trees F and G are equally complex in terms of their graph-based measure. Both are also played against the same opponent, Bob. At the same time, for Ann, Tree F is significantly simpler: she does not have to reason at all because she wins no matter what she does. Naturally, payoff distribution can make a tree less or more complex for a human; here, we can ponder the concept of payoff-based complexity.

[Figure 3 about here.]

To sum up, the profiles of opponents as well as payoff distribution are, without a doubt, important factors determining how complex a game is. One can come up with other variables affecting Ann's perception of complexity. For each, there is a need to start with theoretical foundations to develop a data-based measure of complexity. Naturally, it is impossible to study all the relevant factors at once as scientific standards require focusing on just one variable at a time.

In this paper, our focus is entirely on the graph as a source of tree complexity. We start with the theoretical notion of graph-based complexity and continue with its empirical extension. To that end, we need to cancel, or at least minimize, the impact of other complexity-relevant factors; in Section 3, we elaborate on the experimental design. We leave it for future research to study opponent-based, payoff-based complexity, and other factors affecting game

complexity.

3 Experimental design

In this section, we describe *Blues and Reds* as an experiment. Since the app is available for free for both iOS and Android devices, we encourage readers to install and play *Blues and Reds*. Experiencing what the subjects had experienced would help the readers understand our experiment better than any words could. Moreover, for the readers unable to play *Blues and Reds*, Appendix B describes the app from the gaming perspective.

Our experiment focuses on finite dynamic games with complete and perfect information, no-tie zero-sum payoff structure, and two players (human subject versus the algorithm). For the subject, winning is possible in each game. However, winning requires following the unique set of actions consistent with backward induction. If, at any round, the subject makes a mistake and deviates from that set, then she certainly loses as the algorithm was programmed to backward induction and exploit the subject's mistakes.

3.1 Games in *Blues and Reds*

This paper is based on data from 27 games in *Blues and Reds*. Each game resembles a classical game-theoretic tree; Figure 4 includes a screenshot from the app with an example of a game.

[Figure 4 about here.]

Each game is governed by the same rules. The subject is always the first to choose an action. They start by choosing which blue bridge the RoboToken (golden sphere) is to cross. Next, in the second round, the algorithm decides their action by choosing a red bridge for the RoboToken to cross. In the next round, it is again the subject who moves; then the algorithm, and so on. The subject always chooses at odd rounds, and the algorithm at even rounds.[5] A game ends when the RoboToken lands on a red or blue node. If it is a blue node, then the subject wins and the algorithm loses. If it is a red node, then the subject loses and the algorithm wins.

[5]To avoid making choices by mistake, each time a subject selects an action she has to confirm her choice. In addition, there is no time limit when making choices.

Winning a game results with the subject gaining a star. (In Figure 4, the subject had won so far 5 stars.) This in-game reward system is based on the standards implemented in the mobile gaming industry. It incentivizes subjects to exert effort on reasoning rather than making thoughtless choices and is enforced in the following way: in each of the 27 games which generate the data this paper is based on, subjects have only one chance to play. Whether they win or lose, they are unable to play the game again.

The algorithm was designed to backward induct. No matter how complex a game is, the algorithm waits 3 seconds before making its move. We opted for the constant response time of the algorithm as a varying response time would suggest a game's complexity level which might influence subject's behavior.

In each game, a subject can win but this requires choosing bridges in accordance with backward induction. Importantly, all the games are characterized by the "no mistake" property: starting from the initial node, there is only one action that would guarantee a success. If, at any node, a subject chooses incorrectly, then the subject's loss is guaranteed as the algorithm was programmed to exploit the subject's mistakes.

Each game in our sample has a symmetrical structure $N_1.N_2.N_3.N_4.N_5.N_6$, where N_i is the number of actions at round i. For simplicity, zeros are omitted. For instance, Figure 4 depicts game 3.2.2.2. Trees A, B, and C in Figure 1 are labeled as 2.2, 3.2, and 2.3, respectively. The sample includes 2-, 3-, 4-, 5-, and 6-round games; all are listed in Table 1. Appendix A provides screenshots of all the 27 games from *Blues and Reds* used in this paper. The order of the games a subject plays is randomly assigned.

Games in *Blues and Reds* allow us to test backward induction: subject's winning a game signifies she correctly backward inducted in that game, and her losing means she did not.[6] Because testing backward induction is a challenge facing well-known difficulties – social preferences and belief that opponents are not fully rational – below, we explain how the design of *Blues and Reds* overcomes these difficulties.

As well-documented in the literature, when social preferences matter (e.g., altruism in McKelvey and Palfrey (1992) or fairness in Güth et al. (1982)), subjects make choices which are seemingly inconsistent with backward induction despite actually correctly backward inducting. In *Blues and Reds*, humans play against the computer which, as argued in Johnson et al. (2002), "turns off social preferences (and beliefs that other players express social pref-

[6]Obviously, we follow the typical "as if" view in economics: a subject is said to correctly backward induct if their choices are consistent with the outcome of implementing backward induction . We do not expect the subject to know and literally implement backward induction.

erences) by having human subjects bargain with robot players who play subgame perfectly and maximize their own earnings, and believe the humans will too." Consequently, the presence of social preferences in *Blues and Reds* is negligible at best and we do not consider social preferences to add a typical noise to the observed behavior.

It also is possible that subjects believe that their opponents are not fully rational (Palacios-Huerta and Volij (2009), Agranov et al. (2012), Alaoui and Penta (2016), Fehr and Huck (2016), and Gill and Prowse (2016)), which also might lead to behavior seemingly inconsistent with backward induction. In *Blues and Reds*, all games are zero-sum games. As Levitt et al. (2011) observe, "behavior in these games does not depend on social preferences or beliefs about the rationality of one's opponent. This allows for a purer measure of players' ability to recognize and implement backward induction strategies." Consequently, what human subjects think about the algorithm's rationality is not of relevant concern.

3.2 Procedure: recruitment and instructions

Subjects were recruited via Google Play and the App Store; responding to the online advertisement (AdWords), word-of-mouth, or promotion via radio and traditional media, people who install and play *Blues and Reds* become subjects of our mobile experiment.

In order to provide experimental instructions, every subject must go through a mandatory tutorial that consists of two games. Appendix B.4 includes the screenshots from the mandatory tutorial.

During the tutorial session, subjects learn that they play the winner-takes-all games against the algorithm. They discover their and the algorithm's utilities (blue and red node meaning a win and loss, respectively, for the subject). They practice making choices and learn that there is no time limit in playing the games.

The tutorial can be repeated at any time and is accompanied with a tutorial-text that also explains the app and how to use it. (The tutorial text is available in Appendix B.4.) While the tutorial teaches subjects how to play and what their goal is, it does not suggest using backward induction nor does it inform that making a mistake guarantees a loss.

4 Finding empirical measure of tree complexity

4.1 Data and sample

We collected data from August 15, 2017 to February 6, 2018. In our analysis, we only use data from subjects who played at least one game of *Blues and Reds* beyond the mandatory tutorial. For each game and each subject, we collected the following variables:

i. Whether a subject wins or loses. As explained in Section 3, we say that a subject who won a game was correctly backward inducting in that game. Those who lost are said to have not correctly backward inducted.

ii. At each round, response time (RT), which measures how many seconds a subject spends on selecting an action. RTs capture the subject's reasoning process. For a given game and subject i, let RTk_i denote the subject's response time at the kth round and let TT_i be the total time the subject spent on reasoning in a given game; i.e., $TT_i = \sum_{k=1}^{M} RTk_i$ where M denotes the number of rounds at which subjects choose actions.

Appendix D presents, for each game, distribution of winners for each quintile of TT and $RT1$.

Given the large size of our data, we take a conservative approach and, for each game, keep only the observations in which the subject's total time (i.e., the sum of RTs across all rounds) is below the 95th percentile of its sample. Our final sample consists of 6,637 subjects who played 44,113 games; on average, a subject participated in 6.65 games.

Subjects represent 143 countries with the ten most popular being Mexico (13.82% of the sample), India (12.14%), Argentina (6.71%), Brazil (6.13%), Colombia (4.84%), USA (4.20%), Spain (3.95%), Chile (3.74%), Poland (3.35%), and Thailand (2.85%).[7]

4.2 Graph-based measure of tree complexity

The sample consists of 27 games, which implies 351 pairs of comparable trees. Out of those 351 pairs, 136 can be ranked by the graph-based measure of complexity. As explained in Section 2, an graph-based less complex tree can be embedded in a more complex tree. For example, 3.3.3 is graph-based more complex than 2.2 because we obtain 3.3.3 from 2.2 by

[7]It was possible to identify subject's location by her device's IP for 5,905 subjects.

adding branches and nodes to 2.2. Formally, game $N_1 \_ .. \_ N_k$ compared to $M_1 \_ .. \_ M_l$ is graph-based more complex if the following holds.

1. If $k = l$, then $N_i \geq M_i$ for each i with at least one inequality being strict. (E.g., 3.3.3 is graph-based more complex than 3.2.3.)

2. If $k > l$, then $N_1 \geq M_1$,..., $N_l \geq M_l$. (E.g., 3.3.3 is graph-based more complex than 3.3.)

Figure 5 depicts all graph-based pairwise comparisons of complexity among the 27 games in our sample. Take a game X from the horizontal axis and a game Y from the vertical axis. If the symbol at the intersection of the Xth column and Yth row is $\trianglerighteq$, then Y is graph-based more complex than X, and if the symbol is a dot, then it is not possible to graph-based compare X and Y. As previously discussed, the graph-based measure of tree complexity is incomplete which is reflected by the dots in Figure 5.

[Figure 5 about here.]

4.3 Empirical measure of tree complexity

The main goal of this paper is to derive the empirical measure of tree complexity that will fill the gaps in Figure 5. To that end, we consider three candidates.

1. $ART1$. Average response time at the first round of a given game. Suppose that N subjects played a given game; then $ART1 = \frac{1}{N} \sum_{i=1}^{N} RT1_i$. We consider this metric for the following reason. Backward induction requires the subject to reason at the beginning of a dynamic game; i.e., the first round. Hence, $RT1_i$ measures i's thinking effort at the most important time of playing a game in *Blues and Reds*. The larger that initial effort the more complex a game is.

2. ATT. Average total time spent on solving a game. Suppose that N subjects played a given game; then $ART1 = \frac{1}{N} \sum_{i=1}^{N} TT_i$. The intuition behind this metric is the following. Total time is a proxy for the total effort a subject spends on solving a game. The larger that total effort the more complex a game is.

3. $\%NOT.BI$. Percentage of subjects who did not backward induct in a given game. This metric measures the performance of subjects and is based on the argument that

the performance is expected to decrease when a game becomes more complex; hence, the percentage of those who incorrectly backward induct should increase with a game's complexity.

Table 1 shows the summary statistics of the three candidate variables for each tree.

[Table 1 about here.]

Each of the candidates offers a complete ranking of trees based on behavioral data. The only differentiating element is how well these candidates extend the graph-based ranking depicted in Figure 5.

To identify the best extension of the graph-based measure, we conduct the following exercise. If game A is graph-based more complex than game B, then we check whether a candidate for the empirical measure of tree complexity ranks game A higher than game B using a one-tailed test of difference in means. If we can reject the null hypothesis that the value of the candidate measure for A is less or equal than its value for B at the 1% level of significance, then we say that the candidate measure agrees with the graph-based measure. If game A is graph-based more complex than game B, but the candidate indicates that B is more complex at the 1% level of significance, then we say that the candidate and the graph-based measure disagree. Finally, if game A is graph-based more complex than game B, but according to the candidate their complexities do not differ at the 1% level of significance, then we say that the result of the comparison is undefined. The results are depicted in Table 2.

[Table 2 about here.]

For $ART1$ and ATT, there are no disagreements with the graph-based ranking. However, there are 9 disagreements between the graph-based ranking and $\%NOT.BI$: 2.2 vs 2.3, 2.2.2.2 vs 2.4.2.2, 2.2.2.2 vs 2.2.3.2, 2.2.2.2 vs 2.2.4.2, 2.2.2.2 vs 2.2.2.3, 2.2.2.2 vs 2.2.2.4, 2.2.2.2 vs 2.2.2.2.2, 2.3.2.2 vs 2.4.2.2, and 2.2.2.3 vs 2.2.2.4.

Looking at Table 2, we conclude that $ART1$ and ATT are very similar and far superior to $\%NOT.BI$. In fact, $ART1$ and ATT almost perfectly replicate the graph-based ranking of complexity. Consequently, we consider both $ART1$ and ATT as the appropriate empirical measures of tree complexity.[8]

[8]Appendix C contains several robustness checks using different subsamples of the data and confirm the results obtained in Table 2: (Appendix C.1) subjects who win a game, (Appendix C.2) subjects who played all 27 games, (Appendix C.3) data trimmed at the 5% and 95% of total time spent solving a game, (Appendix C.4) subjects separated by gender, and the subsample using only the first game played by each subject (Appendix C.5).

While ATT and $ART1$ perform equally well as measures of tree complexity, it is possible to advance the argument that $ART1$ outclasses ATT. First, from the data perspective, $ART1$ is less demanding than ATT as the former requires to measure RTs only at the first round.

Second, and more importantly, ATT, by its very nature of being the sum of RTs from all rounds of a tree, will tend to be larger for trees with a larger number of rounds. By design, ATT is affected by the tree length. This is not the case of $ART1$, an argument in favor of choosing $ART1$ over ATT.

The failure of $\%NOT.BI$ as a measure of complexity is not entirely surprising. Recall that, in this paper, we think of complexity as indicative of difficulty. We believe that, in general, the percentage of people who succeed at a task does not necessarily indicate how easy or hard the task is. This is especially true for tasks which are (i) extremely difficult/easy or (ii) similar but different. Below, we elaborate.

First, suppose that we want to measure whether climbing Mount Everest is more difficult than climbing K2. For an average person, it is fair to say that each task is rather extreme. It also is reasonable to expect that in a sample consisting of random people, the percentage of those who succeed at either task is precisely zero indicating that both mountains are equally challenging. Yet, we know from the experts that this is not the case as K2 is considered significantly more difficult.[9]

Second, consider two similar but different tasks. Let task A consist of solving the equation $1 + 3$ while task B be about solving the equation $4 + 1 + 3$. It is likely that a random subject either solves both tasks or fails at both of them. By looking only at the percentage of winners and losers, we would be unable to differentiate between tasks A and B. At the same time, we feel that these tasks are not equally complex. For instance, for a child who just started learning addition, task A might be doable while task B too difficult. Yet, children who allow to distinguish between the two tasks are unlikely to be a significant part of our sample making conclusions based on the percentage of winners/loser potentially erroneous.

It is the second concern that is of particular relevance in our experiment. As discussed above, $\%NOT.BI$ disagrees with the graph-based measure in 9 pairwise comparisons. Importantly, in each pair, the ranked trees are relatively similar. At the end, $\%NOT.BI$ proves to be too crude of a measure. On the other hand, a measure based on response times offers a more refined comparison.

[9]Looking at the percentages generated by a non-random samples is problematic as well due to the self-selection bias. Only the best mountaineers attempt to climb K2. Consequently, comparison of success rates between Mount Everest and K2 is meaningless.

4.4 Tree complexity and backward induction

Table 1 (mean of $\%NOT.BI$) shows that between 2% and 53% of subjects fail to backward induct. Our data not only confirms what is already known in the literature – namely, people do not backward induct – but also points at tree complexity as a new explanation of why backward induction is being violated.

In Table 1, we observe that $\%NOT.BI$ is not the same across the 27 games in our sample. To explain this variability, in Figure 6, we plot the empirical measure of tree complexity ($ART1$) on the horizontal axis and the failure rate of backward induction ($\%NOT.BI$) on the vertical axis. We observe that the higher the complexity the more likely our subjects are to behave in disagreement with backward induction.

[Figure 6 about here.]

A reader might consider this result as trivial, which indicates it is important to report it because, given that it is precisely what one would expect, it validates the proposed measure of tree complexity.

5 Tree structure and complexity

In this section, our objective is to analyze the main drivers of empirical complexity. In particular, we look for connections between the tree structure and its complexity. In what follows, we use $ART1$ as the measure of complexity.

Starting with a given tree, we can add more rounds (make the tree longer) or add more actions at the already existing nodes (make the tree wider). The three questions we ask are the following:

1. Is a longer tree more complex?
2. Is a wider tree more complex?
3. What increases the complexity of a tree more: length or width?

First, we analyze the impact of length. To that end, we group the trees by their number of rounds and, for each group, calculate its average complexity, that is, the weighted average (by the number of subjects) of the empirical complexity of trees in a group. Figure 7 depicts

the average complexity by the number of rounds. To make the comparison visually simple, we normalize the complexity measure to have the value of 1 for the group of trees with two rounds. Figure 7 clearly shows that increasing the number of rounds increases the empirical complexity of a tree.[10]

[Figure 7 about here.]

We now study the impact of increasing the number of actions per round. The analysis is presented in Table 3.

[Table 3 about here.]

First, we partition trees according to the number of rounds. Second, we divide each subset of trees with the same number of rounds into three groups — Min, Med, and Max — according to the number of actions per round. This is captured in Panel (a). Finally, for each group of trees, we compute the average empirical complexity; this is captured in Panel (b).

To understand the impact of the width on tree complexity, we look at Panel (b) in Table 3 column by column. For a fixed number of rounds, we say that the tree becomes wider when we go from Min to Med to Max rows. We observe that for all number of rounds, making a tree wider makes it more complex.

Finally, we analyze what has a bigger impact on the complexity of a tree: making it longer or making it wider. To that end, we propose the following exercise. We take a tree X and expand it in two directions: make it longer (X_L) and make it wider (X_W) but with a constraint that the number of final nodes in X_L and X_W is the same.[11] Next, we compare the complexities of X_L and X_W. If the former is more complex than the latter, then we say that the length is relatively more important than the width; otherwise, we say that it is the width that is more important.

In our data, there are four cases of trees that we can make longer and wider while keeping the number of final nodes the same after each expansion. In Figure 8, we graphically depict

[10]Note that the increment from the 5th to the 6th round looks small. This is because the group of 5-round games contains three games, 2.2.2.2.2, 3.2.2.2.2, and 4.2.2.2.2, while the group of 6-round games contains only 2.2.2.2.2.2. However, if we just increase the number of rounds keeping the number of actions constant from 2.2.2.2.2 to 2.2.2.2.2.2, the measure of complexity increases by 30%, from 43.33 to 56.16.

[11]This constraint is important as the number of final nodes represents the number of paths that a decision-maker must analyze. Without our constraint, we could elongate and widen the initial tree X in any way, which would render the analysis very hard to interpret.

our analysis and discuss it in detail below.[12]

[Figure 8 about here.]

Case 1. We start with the tree 2.2 (4 final nodes), whose empirical complexity is 9.09. We can make 2.2 longer by expanding it to 2.2.2 (8 final nodes); with this, the empirical complexity increases to 13.53. Alternatively, we can make 2.2 wider by expanding it to 2.4 (8 final nodes); the empirical complexity increases to 9.53. We observe that elongation is more important.

Case 2. We start with the tree 2.2.2 (8 final nodes), whose empirical complexity is 13.53. We can make 2.2.2 longer by expanding it to 2.2.2.2 (16 final nodes) with the empirical complexity 21.61. Alternatively, we can make 2.2.2 wider by expanding it to 4.2.2 (16 final nodes) with the empirical complexity 17.31. We observe that elongation is more important.

Case 3. We start with the tree 2.2.2.2 (16 final nodes), whose empirical complexity is 21.61. We can make 2.2.2.2 longer by expanding it to 2.2.2.2.2 (32 final nodes) with the empirical complexity 43.33. Alternatively, we can make 2.2.2.2 wider by expanding it to one of the following trees: 2.4.2.2, 2.2.4.2, 2.2.2.4, or 4.2.2.2. Each of these trees has 32 final nodes. Their average complexity is 29.53. We observe that elongation is more important.

Case 4. We start with the tree 2.2.2.2.2 (32 final nodes), whose empirical complexity is 43.33. We can make 2.2.2.2.2 longer by expanding it to 2.2.2.2.2.2 (64 final nodes) with the empirical complexity 56.16. Alternatively, we can make 2.2.2.2.2 wider by expanding it to 4.2.2.2.2 (64 final nodes) with the empirical complexity 72.27. We observe that elongation is less important.

We note that in 3 out of 4 cases, it is the length that has a bigger impact on complexity.

To summarize our analysis of length and width, we answer the three questions we asked as follows.

1. Is a longer tree more complex? Yes.

2. Is a wider tree more complex? Yes.

3. What increases the complexity of a tree more: length or width? Length.

[12]The difference among $ART1$ values connected by the arrows in Figure 8 is statistically significant at the 1% level or less using a standard one-tailed test of difference in means.

6 Controlling for complexity: omitted-variable bias problem

Section 5 shows the benefits of having an empirical measure of tree complexity. In this section, we focus on the potential consequences of not having such a measure. We provide an example in which we try to understand the impact of experience on the probability of a subject correctly solving a tree but, at the same time, we neglect to include a complexity measure. With this mistake, we encounter the omitted-variable bias problem.

First, we define the variable $Seq = 1, ..., 27$ as the order in which a game has appeared to a subject. Remember that the sequence in which the games appear is randomized for each subject and goes from 1 to 27. Now let's assume that an econometrician has access to a subset of our data that is large enough to test the impact of experience on the probability of correctly solving a dynamic game. This subset of the data consists of just six games and is shown in Table 4, where three of the six games have (randomly) appeared to subjects in one of the first three positions in their sequence ($Seq \leq 3$) and the other three of the six games have (randomly) appeared after the third game played ($Seq > 3$).

We claim that Seq captures experience because the larger Seq is, the more games the subject has played. Table 4 also shows the number of subjects in this subsample and the empirical measure of complexity previously calculated (see Table 1).[13]

[Table 4 about here.]

Armed with this database of 4,852 observations, our hypothetical econometrician decides to test how experience, captured by the variable Seq, affects the probability of correctly solving a game. For this purpose, the econometrician estimates the following logit regression:

$$logit(Y_{i,Seq}) = \alpha + \beta_S Seq_i \tag{1}$$

Y_i is the dependent variable in the regression and captures whether the subject i correctly solved the game appearing in the order Seq ($Y_i = 1$) or did not ($Y_i = 0$), α is the intercept, and Seq_i corresponds to the order in which a game appeared in the subject i's sequence of

[13] For an analysis of the impact of experience on the probability of a subject correctly solving a game using the entire sample please see Appendix E. Differently to the example provided in this section with a modified subsample of the data, Appendix E shows that the randomization of the sequence of games implemented in *Blues and Reds* effectively removes the impact of subjects' experience on complexity. It also shows that experience has a positive impact on the probability of a subject correctly solving a game.

games. Results from this regression are shown in Column (1) of Table 5 below.

[Table 5 about here.]

In this case, the estimated parameter of Seq (β_S) is negative and statistically significant, leading the econometrician to conclude that experience is detrimental to the probability of a subject correctly solving a game. As we show below, this unexpected result disappears once we run a correct model in which we control for complexity using a measure proposed in this paper. Column (2) of Table 5 shows the results of estimating the following model:

$$logit(Y_{i,Seq}) = \alpha + \beta_S Seq_i + \beta_A ART1_i \tag{2}$$

where $ART1_i$ is the empirical measure of complexity for a game appearing in the order Seq of subject i.

The results from regression (2) show that now the coefficient of Seq is positive and statistically significant, leading the econometrician to correctly conclude that experience improves the probability of a subject correctly solving a game. The negative and significant coefficient of $ART1$ shows that more complex games are less likely to be correctly solved.

Adding the empirical measure of complexity $ART1$ solved the omitted-variable bias problem in this example, showing its potential use in many other situations in which a researcher needs to control for complexity.[14]

7 Conclusions

When adding numbers, not all problems are equally complex; e.g., 2 + 2 is less complex than 17 + 23 + 19. The same reasoning applies to backward induction: trees differ in their complexity. This paper is devoted to empirically deriving the measure of tree complexity starting with the notion of an graph-based measure of tree complexity: a tree X is graph-based less complex than tree Y if Y can be constructed by adding branches and nodes to X. Since the graph-based measure is incomplete, we propose the notion of an empirical measure of tree complexity defined as a complete and best extension of the graph-based measure based on data generated by subjects playing the trees.

[14]The omitted-variable bias problem in our example manifests in a downward bias estimate of β_S in equation (1) because the correlation between Seq and $ART1$ is positive by construction of this hypothetical database (correlation coefficient equal to 0.63) while $\beta_A < 0$ as shown in Table 5.

Of course, graph-based perspective is not the only one we can take to look at the problem of complexity. Who the opponents are and the payoff distribution are natural factors to consider when studying complexity. Opponent's rationality can make a game less complex (e.g., playing against someone who is known to be perfectly rational vs someone whose rationality is uncertain). A specific payoff structure can make a game trivial (e.g., a chess game with modified rules: no matter what happens, White always wins).

Given our focus on the graph-based perspective, our article's limitations need to be mentioned. The measure of empirical complexity that we advance need not apply to games in which humans play against humans or with a payoff structure more sophisticated than the one we used in our experiment. Measuring opponent-based and payoff-based complexity is not a question that we can address in this article. Nor is the question of how various complexity-relevant factors affect the overall complexity. Hopefully, these are the questions to be addressed in future research. This article makes a first step towards understanding complexity but, indeed, not the final step.

To collect data, we conducted a mobile experiment. To that end, we recruited professional app developers to create *Blues and Reds*, a mobile app that functions as experiment and generates data. The app includes interactive 27 games played against the algorithm. These games generate the data used in this paper.

For each game, *Blues and Reds* records whether a subject wins or loses and her response times. Using this data, we propose three candidates for the empirical measure of tree complexity. Comparing these candidates to the graph-based measure of tree complexity, we select the average response time that subjects spend at the first round of a tree ($ART1$) and the average total time that subjects spend solving the tree (ATT) as the empirical measure of tree complexity. Yet, we also advance arguments favoring $ART1$ over ATT: the latter is more data-demanding and, by definition, larger in trees with a larger number of rounds.

By highlighting tree complexity as a reason for subjects to deviate from the behavior prescribed by backward induction, this paper contributes to the literature studying violations of backward induction. By deriving an empirical measure of tree complexity, this paper provides a control variable for studies in which it is necessary to empirically control for the complexity of dynamic games that subjects play.

We employ the proposed complexity measure for two empirical exercises. Investigating the relationship between a structure of a tree and its complexity, we find that both making a tree longer and making a tree wider increases its complexity but it is the length that has a larger impact. We also show that not controlling for complexity is potentially of serious

concern as it might lead to the omitted-variable bias problem.

Our analysis is focused on two-person, finite games with perfect and complete information and no ties. Naturally, one wonders whether the results would extend to other games, especially those with more than two players or non-zero-sum payoffs. This remains to be answered in future research.

References

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.

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.

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.

Caplin, A. and D. Martin (2016): "The Dual-Process Drift Diffusion Model: Evidence from Response Times," *Economic Inquiry*, 54, 1274–1282.

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. (2018): "Response Times in Economics: Looking Through the Lens of Sequential Sampling Models," *Journal of Economic Psychology*, 69, 61–86.

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein (2009): *Introduction to Algorithms*, MIT Press.

Enke, B. and F. Zimmermann (2019): "Correlation Neglect in Belief Formation," *Review of Economic Studies*, 86, 313–332.

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

Fehr, E. and A. Rangel (2011): "Neuroeconomic Foundations of Economic Choice–Recent Advances," *Journal of Economic Perspectives*, 25, 3–30.

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*.

Grabiszewski, K. and A. Horenstein (2020a): "Effort is not a monotonic function of skills: results from a global mobile experiment," *Journal of Economic Behavior & Organization*, 176.

——— (2020b): "Game-form recognition in dynamic interactions," *working paper*.

Güth, W., R. Schmittberger, and B. Schwarze (1982): "An experimental analysis of ultimatum bargaining," *Journal of Economic Behavior & Organization*, 3, 367–388.

Johnson, E. J., C. Camerer, S. Sen, and T. Rymon (2002): "Detecting Failures of Backward Induction: Monitoring Information Search in Sequential Bargaining," *Journal of Economic Theory*, 104, 16–47.

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.

Krajbich, I., B. Oud, and E. Fehr (2014): "Benefits of Neuroeconomic Modeling: New Policy Interventions and Predictors of Preference," *American Economic Review*, 104, 501–506.

Levitt, S. D., J. A. List, and S. E. Sadoff (2011): "Checkmate: Exploring Backward Induction among Chess Players," *American Economic Review*, 101, 975–990.

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.

McKelvey, R. D. and T. R. Palfrey (1992): "An Experimental Study of the Centipede Game," *Econometrica*, 60, 803–836.

Ofek, E., M. Yildiz, and E. Haruvy (2007): "The Impact of Prior Decisions on Subsequent Valuations in a Costly Contemplation Model," *Management Science*, 53, 1217–1233.

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

Recalde, M. P., A. Riedl, and L. Vesterlund (2018): "Error-prone inference from response time: The case of intuitive generosity in public-good games," *Journal of Public Economics*, 160, 132–147.

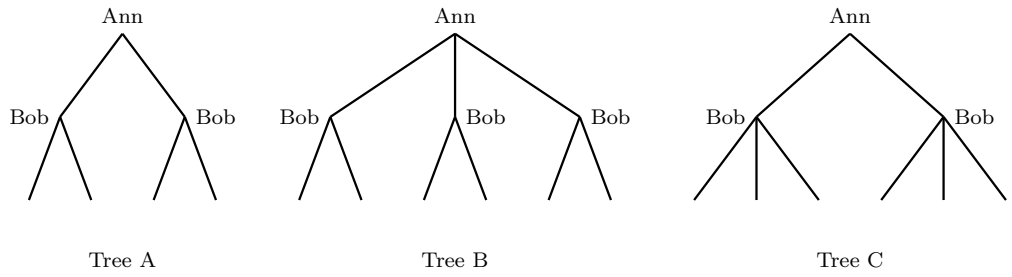
Rubinstein, A. (2007): "Instinctive and Cognitive Reasoning: A Study of Response Times," *Economic Journal*, 117, 1243–1259.

——— (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.

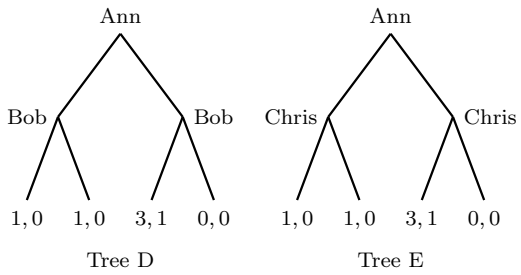
Spiliopoulos, L. and A. Ortmann (2017): "The BCD of Response Time Analysis in Experimental Economics," *Experimental Economics*, 47, 1–55.

Figure 1: Ranking trees by their graph-based complexity.



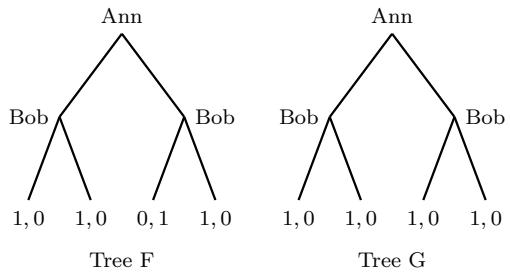
Notes. In terms of their graphs, Tree A is less complex than both Trees B and C. However, graph-based complexity is unable to rank Trees B and C.

Figure 2: Complexity and opponents.



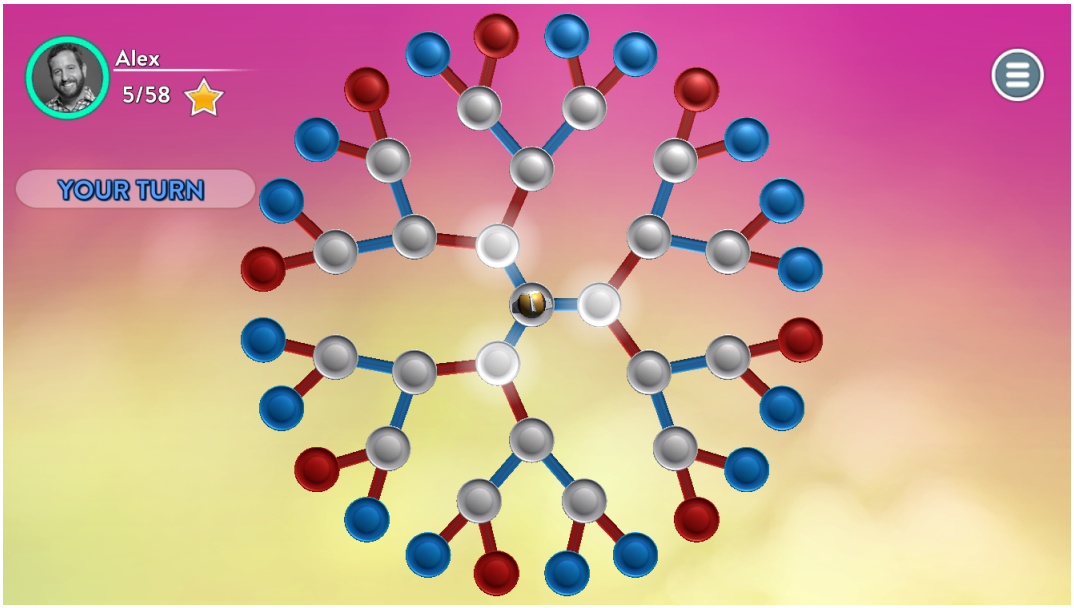
Notes. In terms of their graphs, Tree D and E are equally complex. However, Ann's belief in rationality of Bob and Chris makes these two games of different complexity from Ann's perspective.

Figure 3: Complexity and payoffs.



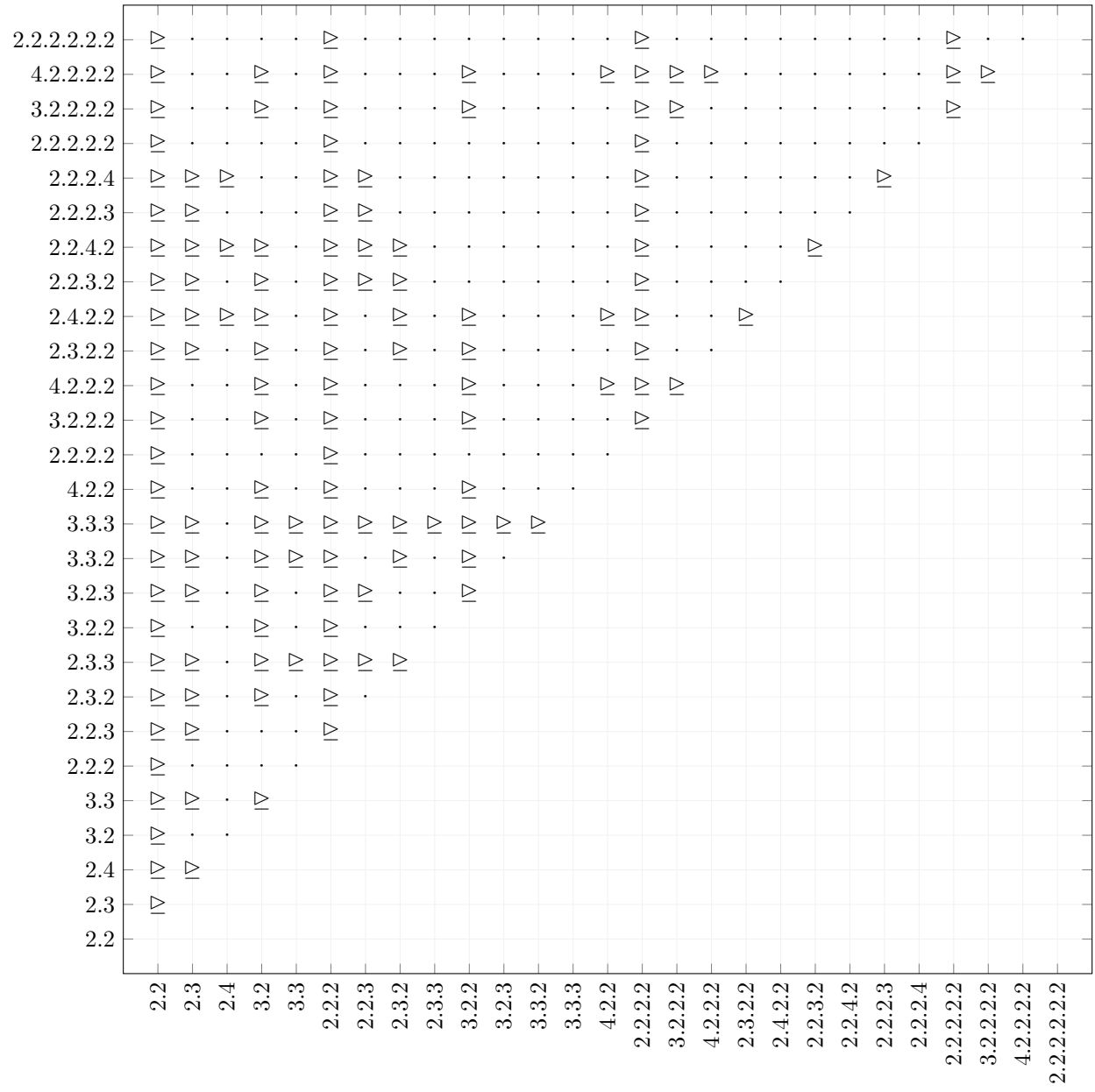
Notes. In terms of their graphs and opponents, Trees F and G are identical from Ann's perspective. However, they differ in the payoff distribution.

Figure 4: A screenshot from *Blues and Reds* with an example of a game.



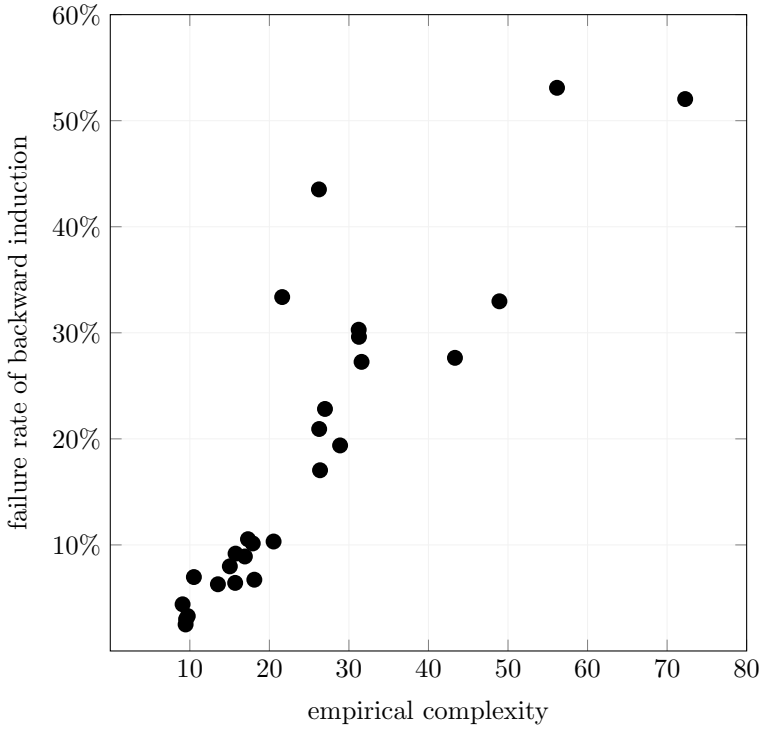
Notes. The subject and algorithm move the golden spherical object (called the RoboToken) across the blue (subject) and red (algorithm) bridges. Choices are made in turns with the subject moving at odd rounds (first, third, etc.). Subject wins if the RoboToken lands on a blue node; otherwise, the subject loses.

Figure 5: Pairwise comparison of games in terms of graph-based tree complexity.



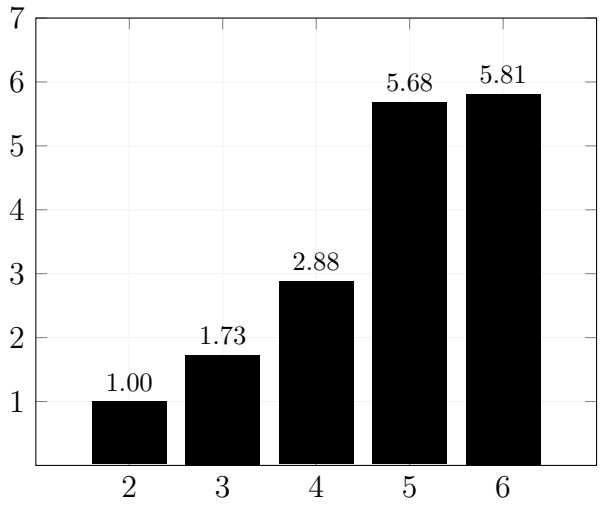
Notes. This figure presents the pairwise comparison of games in accordance with their graph-based complexity. Take a game X from the horizontal axis and a game Y from the vertical axis. If the symbol at the intersection of the Xth column and Yth row is $\trianglerighteq$, then Y is graph-based more complex than X, and if the symbol is a dot, then it is not possible to compare X and Y using graph-based complexity.

Figure 6: Empirical complexity and failure rate of backward induction.



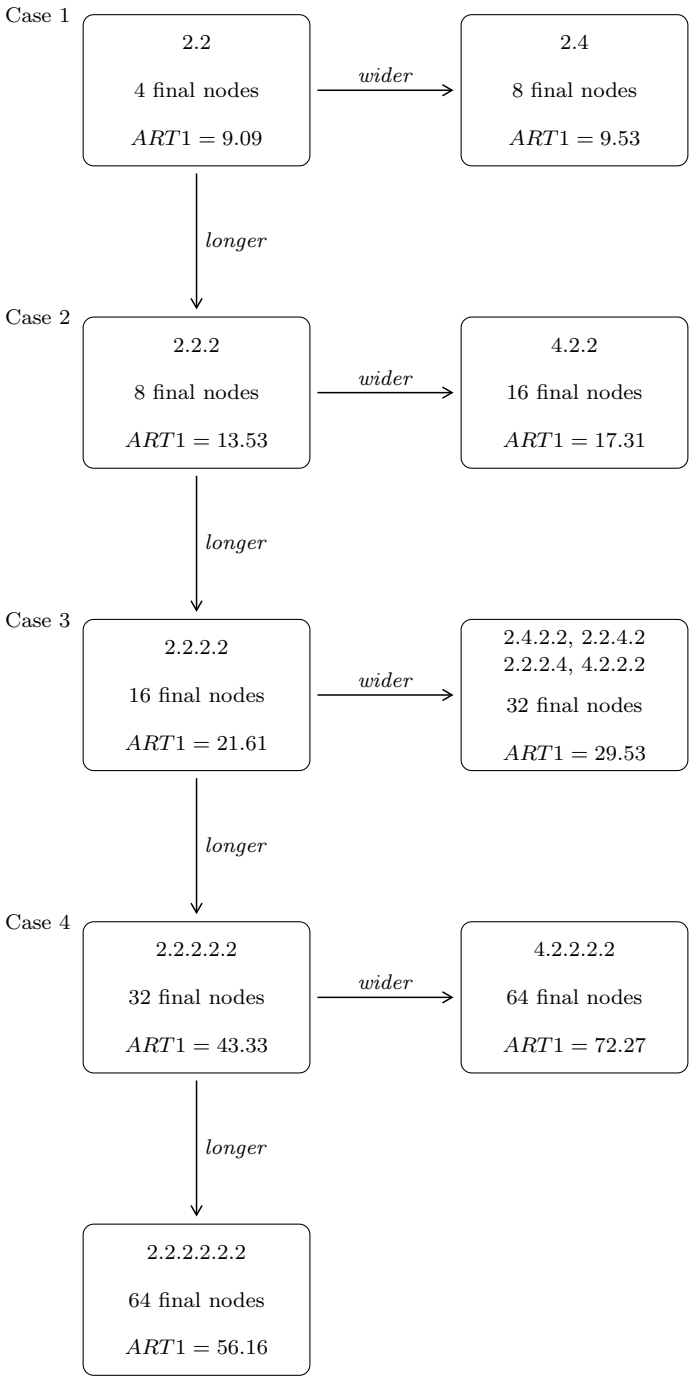
Notes. This figure presents for each of the 27 games in *Blues and Reds*, their values of average response time at the first round ($ART1$) in the horizontal axis with their corresponding value of percentage of subjects who did not backward induct ($\%NOT.BI$) in the vertical axis.

Figure 7: Empirical complexity and length.



Notes. This figure presents the weighted average of empirical complexity in trees with 2, 3, 4, 5, and 6 rounds. Recall that the empirical complexity is measured as $ART1$ (average response time at the first round). We normalize the complexity measure to have the value of 1 for the group of trees with two rounds.

Figure 8: Length vs width analysis.



Notes. This figure presents the analysis of four cases in which we compare the relative impact of length and width on the complexity of trees. In each case, we start with a tree, which we expand in two ways: making it wider and making it longer; however, in each case of expansion, we keep the same number of final nodes. The comparison of length versus width consists of comparison of the complexity of the tree, measured as $ART1$, that results from making the benchmark tree longer and wider.

Table 1: Summary statistics.

Game	N	$ART1$					ATT					$\%NOT.BI$			
		Mean	SD	Med	Max	Min	Mean	SD	Med	Max	Min	Mean	SD	Max	Min
2.2	1,683	9.09	3.13	8	23	3	9.09	3.13	8	23	3	0.04	0.21	1	0
2.3	1,681	9.47	3.36	9	23	4	9.47	3.36	9	23	4	0.02	0.16	1	0
2.4	1,632	9.53	3.34	9	23	2	9.53	3.34	9	23	2	0.03	0.17	1	0
3.2	1,670	9.74	3.64	9	25	3	9.74	3.64	9	25	3	0.03	0.18	1	0
3.3	1,621	10.51	3.64	10	25	2	10.51	3.64	10	25	2	0.07	0.25	1	0
2.2.2	1,638	13.53	6.01	12	45	2	19.05	7.17	17	50	5	0.06	0.24	1	0
2.3.2	1,630	15.03	6.53	13	45	4	21.04	7.86	19	50	8	0.08	0.27	1	0
2.2.3	1,729	15.7	6.36	14	45	2	21.07	7.22	19	49	6	0.06	0.25	1	0
3.2.2	1,666	15.75	6.71	14	45	4	21.03	7.76	19	49	7	0.09	0.29	1	0
3.2.3	1,628	16.92	7.45	15	51	3	21.94	8.44	20	56	8	0.09	0.28	1	0
4.2.2	1,717	17.31	8.2	15	59	3	22.68	9.79	20	62	6	0.11	0.31	1	0
3.3.2	1,647	17.91	8.27	16	57	4	23.18	9.56	21	61	8	0.1	0.3	1	0
2.3.3	1,637	18.11	8.78	15	58	3	22.98	9.82	20	62	8	0.07	0.25	1	0
3.3.3	1,638	20.53	8.77	18	54	2	25.64	9.63	23	58	5	0.1	0.3	1	0
2.2.2.2	1,660	21.61	12.71	18	81	3	30.1	14.6	26	85	6	0.33	0.47	1	0
2.3.2.2	1,606	26.23	16.85	21	97	3	35.98	18.44	31	101	6	0.44	0.5	1	0
2.2.3.2	1,674	26.99	15.67	23	86	4	34.55	16.91	30	90	9	0.23	0.42	1	0
2.2.2.3	1,610	26.26	15.67	21	88	3	33.24	17.01	28	93	6	0.21	0.41	1	0
3.2.2.2	1,575	31.23	19.74	26	115	2	38.44	20.65	33	118	5	0.3	0.46	1	0
4.2.2.2	1,614	31.26	20.44	25	113	5	38.63	21.69	32	121	8	0.3	0.46	1	0
2.4.2.2	1,641	28.89	17.28	24	92	4	35.84	17.94	31	96	9	0.19	0.4	1	0
2.2.4.2	1,673	31.59	19.07	26	108	1	40.44	20.88	35	112	7	0.27	0.45	1	0
2.2.2.4	1,602	26.37	15.16	22	81	3	32.57	15.96	29	87	7	0.17	0.38	1	0
2.2.2.2.2	1,545	43.33	30.45	36	172	2	59.43	32.15	52	184	10	0.28	0.45	1	0
3.2.2.2.2	1,550	48.92	41.1	36	225	2	66.26	42.55	54	235	12	0.33	0.47	1	0
4.2.2.2.2	1,566	72.27	74.78	45	509	2	95.08	78.1	70	534	10	0.52	0.5	1	0
2.2.2.2.2.2	1,580	56.16	55.34	35	311	2	81.55	60.14	62	328	11	0.53	0.5	1	0

Notes. For each game, we present the number of subjects and summary statistics (mean value, standard deviation, median, maximum value, and minimum value) of the three candidate measures of complexity: average response time at the first round ($ART1$), average total time (ATT), and percentage of subjects who did not backward induct ($\%NOT.BI$).

Table 2: Selecting the empirical measure of tree complexity.

	Agree	Disagree	Undefined
$ART1$	133	0	3
ATT	132	0	4
$\%NOT.BI$	105	9	22

Notes. We compare each candidate for the empirical measure of complexity to the graph-based measure of complexity. If game A is more complex than game B, then we check whether a candidate for the empirical measure of complexity ranks game A higher than game B at the 1% level of significance using a one-side t-statistic. If this is the case, then we say that a candidate for the empirical measure agrees with the graph-based measure. If game A is more complex than game B, but the candidate for the empirical measure indicates that B is more complex at the 1% level of significance, then we say that the candidate for the empirical measure and the graph-based measure disagree. Finally, if game A is more complex than game B, but according to the candidate for the empirical measure their complexities do not differ at the 1% level of significance, then we say that the result of the comparison is undefined.

Table 3: Empirical complexity and width.

Panel (a)

	2 rounds	3 rounds	4 rounds	5 rounds
Min	2.2	2.2.2	2.2.2.2	2.2.2.2.2
Med	2.3 3.2	2.3.2 3.2.2 2.2.3	3.2.2.2 2.3.2.2 2.2.3.2 2.2.2.3	3.2.2.2.2
Max	3.3 2.4	3.3.2 2.3.3 3.2.3 4.2.2 3.3.3	2.2.2.4 4.2.2.2 2.4.2.2 2.2.4.2	4.2.2.2.2

Panel (b)

	2 rounds	3 rounds	4 rounds	5 rounds
Min	9.09	13.53	21.61	43.33
Med	9.61	15.50	27.65	48.92
Max	10.02	18.15	29.55	72.27

Notes. These panels present the analysis of how width influences empirical complexity. For each cell in Panel (a), there is an associated cell in Panel (b) with the average empirical complexity (weighted by the number of subjects in the corresponding trees). Recall that the empirical complexity is measured as $ART1$ (average response time at the first node).

Table 4: Subsample of data.

game	N	$ART1$	Seq
4.2.2.2.2	979	72.27	4 and above
2.2.2.2.2.2	977	56.16	4 and above
2.2.2.2.2	951	43.33	4 and above
2.2.2.3	619	26.26	1 to 3
2.3.3	652	18.11	1 to 3
3.2.2	674	15.75	1 to 3

Notes. List of 6 games used to create a subsample. For each of the six games, we provide the number of subjects who played the game N the game's complexity $ART1$, and the order in the sequence Seq in which the games in this subsample appeared (either $Seq \leq 3$ or $Seq > 3$).

Table 5: Results from the logit regressions.

	(1)	(2)
Seq	-0.02374	0.0414
	(0.0039)	(0.0056)
$ART1$		-0.0397
		(0.0020)
$Pseudo - R^2$	0.01	0.08

Notes. This table shows the result for two logit regressions. In each regression, the dependent variable is Y_i (equals 1 if subject i backward inducts and 0 otherwise). In the first regression (Column (1)), the independent variable is Sequence (Seq, or the order in which a tree was assigned to a subject, $Seq = 1, ..., 27$). In the second regression (Column (2)), the independent variables are Sequence and Complexity ($ART1$, or the average response time at the first round in a given tree). In each regression, the number of observations is 4,852. Robust standard errors are in parenthesis.