# Tree Construction and Backward Induction: A Mobile Experiment[*]

Konrad Grabiszewski[†] Alex Horenstein [‡]

December 2018

## Abstract

Dynamic game theory has two fundamental goals: modeling how decision-makers perceive the interaction they participate in (*tree construction*) and identifying the strategies they select (*backward induction*). In our experiment, subjects often violate the theory (i.e., they do not construct a tree or backward induct), which is because either their skills are too low or the interaction is too complex. We use response times to measure skills and complexity. We find that complexity of interaction increases with its length and width; however, the length has a relatively bigger impact. Improving skills or decreasing complexity increases the likelihood of subjects behaving according to the theory; however, improving skills has a bigger impact. To collect the data, we developed a mobile application, *Blues and Reds*, comprising of 58 dynamic games. Our samples include 4,582 (analysis of tree construction) and 6,677 (analysis of backward induction) subjects coming from over 100 countries.

Keywords: game theory; mobile experiment; tree construction; backward induction

[*] 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, Kfir Eliaz, 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 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; alexhorenstein@gmail.com

# 1 Introduction

In this paper, we experimentally study two fundamental concepts of dynamic game theory: tree construction and backward induction. To construct a tree means to behave as if simplifying a real-life interaction by translating it into a tree. To backward induct means to behave as if solving a tree by applying the backward induction algorithm on that tree. We start with asking whether people construct trees and backward induct. Then, we analyze what causes people to fail at constructing trees and at backward inducting. Finally, we estimate what increases the likelihood of observing people constructing a tree and backward inducting.[1]

**Do people construct trees and backward induct?** Our subjects often fail to construct trees and to backward induct, rendering their behavior often inconsistent with game theory. This confirms what is already established in the literature: in general, people do not seem to be able to correctly understand the game (static or dynamic) they play[2] nor do they choose strategies in accordance with theoretical predictions.[3]

**Why don't people construct trees and backward induct?** When our subjects fail to behave in accordance with game theory, it is because either their skills are too low or the interaction they participate in is too complex.

*Skills.* A priori, it is rather obvious that people are not equally skillful at analyzing interactions (i.e., constructing trees) and choosing strategies (i.e., backward inducting). What is less obvious is how to define and measure "skills."

[1] Tree construction and backward induction are two separate game-theoretic concepts that serve two separate roles. The former is about modeling a real-life interaction and capturing how decision-makers perceive reality. The latter is about solving the tree and indicating what strategies decision-makers select. Consequently, it would seem natural to study tree construction and backward induction in two separate papers. However, at the same time, they complement each other and together represent a complete process of game-theoretic analysis: from a model of interaction (tree construction) to selection of strategies (backward induction). Since our ambition is to offer a comprehensive empirical analysis of dynamic game theory, we merge two studies in one paper.

[2] E.g., Schotter et al. (1994), Rapoport (1997), McCabe et al. (2000), Rydval et al. (2009), Cox and James (2012), Halevy et al. (2012), and Cason and Plott (2014).

[3] The experimental literature on backward induction, not to mention Nash equilibrium, is too large for us to properly review here. The relevant question is not anymore whether people backward induct but why they do not backward induct. Later, we discuss how the literature explains the violations of backward induction.

We say that, when it comes to tree construction (backward induction), Ann has higher skills than Bob if the probability of observing Ann constructing a tree (backward inducting) is higher than that of Bob. We treat skills for tree construction and skills for backward induction as two separate concepts.

We measure a subject's skills by looking at her response times (RTs); i.e., time (in seconds) she spends making a decision. As far as we know, ours is the first paper measuring RTs at each round of a dynamic interaction.[4] RTs reflect a subject's skills because they are the outcomes of choosing how much time to spend at different rounds of a dynamic game.

While we analyze the skills for tree construction and the skills for backward induction separately, we find that the key measure of skills is the same in each case: it is the relative response time at the first round, $RRT1$, defined as the response time at the first round, $RT1$, as the percentage of total time, $TT$, spent on solving a game.

The higher $RRT1$, the more likely we are to observe a behavior consistent with what theory predicts. Interestingly, it is only conditional on $RRT1$ that $TT$ is indicative of skills: lower $TT$ implies that a subject is better at tree construction and at backward induction.[5]

*Complexity.* A priori, it is clear that subjects are more likely to behave in accordance with game theory for simpler interactions; for instance, more people will behave as if constructing a tree and backward inducting in a tic-tac-toe game compared to the game of chess. However, as in the case of skills, there are two challenges with the concept of "complexity:" definition and measure.

We define complexity as an empirical concept. To that end, we begin with the *objective* measure of complexity. We say that an interaction A is objectively less complex than interaction B if the tree representing B can be obtained by extending (i.e., adding

[4] RT is a tool that just 15 years ago was rather unlikely to be seen in economics papers. We believe that the change in the economists' attitudes towards RT is due to the 2004 Presidential Address to the Econometric Society by Ariel Rubinstein (Rubinstein (2006)). Since then, the empirical literature using RT has been growing very fast (e.g., Rubinstein (2007), Rand et al. (2012), Rubinstein (2013), Agranov et al. (2015), Evans et al. (2015), Rubinstein (2016), Gill and Prowse (2017), and Lohse et al. (2017)). For the reviews of the relevant literature, see Clithero (2016) and Spiliopoulos and Ortmann (2017).

[5] If we think about time as a scarce and costly resource, then the allocation of resource (*how* a subject spends time) is of primary importance while the amount of resource (*how much* of time a subject spends) is of secondary importance.

nodes and branches) the tree representing A. In the context of tree construction, being more complex means that an interaction is more difficult to depict as a tree. In the context of backward induction, being more complex means that an interaction is more difficult to solve.

The objective measure of complexity generates the incomplete ranking of interactions. This incompleteness motivates our concept of the *empirical* measure of complexity defined as the best complete extension of the objective measure of complexity. We treat empirical measure of complexity for tree construction and empirical measure of complexity for backward induction as two separate concepts.

We find that the empirical complexity in the context of tree construction as well as in the context of backward induction can be measured in the same way. In fact, there are two measures that do an equally good job: the average response time that subjects spend at the beginning of the interaction (i.e., average $RT1$) and the average total time that subjects spend solving the interaction (i.e., average $TT$).

Our measures of empirical complexity show that making an interaction longer (more rounds) or wider (more available actions) increases the complexity of the game in our studies of backward induction and tree construction. Additionally, we find that increasing the length is more detrimental to a subject than increasing the width.

According to the experimental literature on tree construction, people do not model reality in the way that game theory does because they often do not search for relevant information about the very interaction they participate in.[6] Our paper is complementary to this literature. As far as we know, ours is the first paper to measure skills and complexity using RTs, demonstrate that low skills (subject) and high complexity (interaction) are the reasons for people not constructing trees, and study how the width and length of interaction affects its complexity.

The experimental literature on backward induction attributes the violations of backward induction to the subject's imperfect strategic reasoning,[7] subject's cognitive

[6] E.g., Costa-Gomes et al. (2001), Johnson et al. (2002), Salmon (2004), Costa-Gomes and Crawford (2006), Knoepfle et al. (2009), Wang et al. (2010), Arieli et al. (2011), Reutskaja et al. (2011), Brocas et al. (2014), and Devetag et al. (2016).

[7] Here, the fundamental model is the level-$k$ model. This model was introduced in Stahl and Wilson (1994), Stahl and Wilson (1995), and Nagel (1995). For the literature review on level-$k$ reasoning, see Crawford et al. (2013). See also the closely related model of cognitive hierarchy developed in Camerer et al. (2004). The level-$k$ model has been extensively tested in the literature

skills,[8] and cognitive skills of the subject's opponents.[9] Similarly to the literature, we argue that backward induction fails because of the subject; however, we offer a new, RT-based look at the concept of skills. In addition, to the best of our knowledge, ours is the first paper to measure complexity using RTs, discover that complexity also explains violations of backward induction, and link length and width of interaction to its complexity. Note that Gill and Prowse (2017) also measure the complexity of interaction and, like us, also rely on RTs; however, their study is in the context of static games.

**Complexity versus skills: what matters more?** In our experiment, when people do not behave in accordance with the theory, it is because either their skills are too low or the interaction they are part of is too complex. In our final exercise, we analyze whether improving skills or lowering complexity yields better results in terms of increasing the likelihood of subjects constructing trees and backward inducting. To the best of our knowledge, our paper is the first to address this issue. We find that for both — tree construction and backward induction — improving skills is the best recipe.

Our skills-versus-complexity analysis is important from the perspective of designing dynamic games to be implemented in the real world. For instance, consider the problem of creating a new regulation that aims at modifying the behavior of interacting agents. Game theory predicts what is going to happen and, based on that prediction, the proposed regulation becomes the reality. However, surprisingly, the observed outcome of regulation is less exciting than what theory predicted.

As our results indicate, this gap between theory and reality is because either the population's skills are too low or the regulation (as an interactive problem) is too

and the main message is that people do indeed struggle with strategic reasoning (e.g., 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)). See also experimental studies related to the level-$k$ model in Kneeland (2015), Bayer and Renou (2016a), and Friedenberg et al. (2017).

[8] E.g., Burks et al. (2009), Burnham et al. (2009), Rydval et al. (2009), Brañas-Garza et al. (2012), Carpenter et al. (2013), Duffy and Smith (2014), Agranov et al. (2015), Allred et al. (2016), Bayer and Renou (2016b), Benito-Ostolaza et al. (2016) Gill and Prowse (2016), Hanaki et al. (2016), and Kiss et al. (2016).

[9] E.g., Palacios-Huerta and Volij (2009), Agranov et al. (2012), Alaoui and Penta (2016), Fehr and Huck (2016), and Gill and Prowse (2016).

complex. In order to improve the outcome of regulation, it is necessary to choose between augmenting skills (e.g., education) and making regulation less complex (e.g., simplification of law). The choice crucially depends on whether skills or complexity matters relatively more. Our paper indicates that education focused on enhancing skills is a better choice.

**Mobile experiment.** In order to collect the data, we designed and implemented a mobile experiment, that is an experiment that gamifies a research question into a mobile game and takes place on the subjects' smartphones and tablets.[10] For the purposes of this paper, we developed *Blues and Reds*, a mobile game available globally for free for iOS and Android devices.[11]

Our mobile experiment consists of 58 two-person, turn-based, zero-sum finite games with perfect and complete information in which a subject plays against Artificial Intelligence (AI). The database we use to study backward induction consists of 6,677 subjects, while the database for the study of tree construction consists of 4,582 subjects. In order to facilitate the discussion of our paper, we would like to invite readers to play *Blues and Reds*.[12]

**Structure of the paper.** The rest of the paper is organized in the following way. In section 2, we present *Blues and Reds* as an experiment. In section 3, we find that our subjects sometimes construct trees and sometimes backward induct. In section 4, we explain why this is the case. We also analyze how the width and length of interaction affects the likelihood of our subjects constructing a tree and backward inducting. In section 5, we analyze whether improving skills or lowering complexity yields better results in terms of increasing the likelihood of our subjects behaving in accordance with the theory. We conclude in section 6.

[10] From the methodological perspective, our paper belongs to the line of research that relies on non-standard methods to gather the data, like newspapers-based experiments (e.g., Bosch-Domènech et al. (2002)) and web-based experiments (e.g., Ariel Rubinstein's `gametheory.tau.ac.il`).

[11] Using a mobile game to study game theory poses a linguistic problem as the meaning of the term "game" depends on context. In order to avoid confusion, we say "mobile game" when referring to *Blues and Reds* and "game" for a game-theoretic game.

[12] While there are many advantages of using mobile technology for experimental research, one of them is particularly evident at this very moment: our readers can play *Blues and Reds* and not only be part of the experiment but also, and more importantly, directly examine our experiment. For those who do not have an opportunity to play our mobile game, we prepared the appendix that describes *Blues and Reds* from the gaming perspective.

# 2 Experimental Design

*Blues and Reds* consists of 58 two-person, turn-based, zero-sum finite games with perfect and complete information in which the subject plays against Artificial Intelligence (AI). The first four games constitute the mandatory tutorial in which the subjects learn the rules and how to make choices; we exclude data from these games from our data set.

Games in *Blues and Reds* are divided into two types: tree games and non-tree games (as in Cox and James (2012)). Tree games, which we also call just trees, are classical game-theoretic trees. Modeling them — in the sense that we use in this paper (i.e., tree construction) — is costless. Non-tree games, which we also call non-trees, are depicted in a more convoluted way. For each non-tree, it is possible to construct its tree representation; however, this requires that subjects exert effort.

An important element of our experimental design is that for each tree game there is an equivalent non-tree game (and vice versa). Equivalency means that a tree game and the tree that is an extensive form of the equivalent non-tree game are identical.

The only difference between a tree game and the equivalent non-tree game is how they look. However, "looks" do not matter for theory and, from the game-theoretic perspective, equivalent games are the same interactive situation as they share the same tree representation. In *Blues and Reds*, we use data from 27 pairs of equivalent games.

In *Blues and Reds*, subjects play against AI because we want to eliminate the impact of social preferences. Using AI is motivated by Johnson et al. (2002), whose experiment also involves human vs computer games. They argue that playing against the computer "turns off social preferences (and beliefs that other players express social preferences) by having human subjects bargain with robot players who play subgame perfectly and maximize their own earnings, and believe the humans will too."

In each game, there are only two possible outcomes: either the subject wins and AI loses or the subject loses and AI wins. In addition, information is perfect and complete. This very simple structure allows us to disregard issues like payoff uncertainty (Zauner (1999)) or the experimenter's misunderstanding of the subject's utility function.

Every game is winnable; that is, a subject can win. However, in order to win, the subject must not make a mistake. In the first round, there is only one action that will lead a subject to win the game. This no-mistake property also holds at the subsequent rounds leading towards a win. If a subject makes a mistake at any round, then she will lose for sure.

The no-mistake property is an important feature of our design. First, it minimizes the impact of luck. One can easily imagine an interactive situation in which no matter what a subject does she always wins. Clearly, winning in this case has nothing to do with constructing a tree or backward inducting.

Second, the no-mistake property removes the (ir)rationality of the opponent as a factor affecting the behavior of our subjects. As expected theoretically and confirmed empirically (see footnote 9), what a subject thinks about the rationality of her opponents affects the subject's behavior. However, in our experiment, it does not matter what the subject thinks about AI's rationality. This is because even if the subject assigns non-zero probability to AI making a mistake, then the subject actually has no reason to choose an action different from the one she would choose if the probability of AI being irrational were zero.

Every tree and non-tree has the structure $N_1.N_2.N_3.N_4.N_5.N_6$, where $N_i$ denotes the number of actions at each node at round $i$. We consider $N_i \in \{2, 3, 4\}$ and, in our notation, we omit final zeros; that is, we write, for instance, 2.2.2 instead of 2.2.2.0.0.0.

Table 1 lists all 27 games, sorted by the number of rounds, that we use to build our data. The sequence in which a subject plays the tree/non-tree games is randomly assigned.

Another important aspect of our mobile game is the "one life per game" feature: except for the tutorial games, a subject can play each game only once. While this feature is very uncommon in mobile games, we introduced it in order to motivate subjects to think rather than mindlessly select their choices. Subjects are informed about the "one life per game" feature in the tutorial as well as reminded about it right before they start a new game.

**Data we collected and its interpretation.** We collected data from August 15, 2017 to February 6, 2018. In our analysis, we only use data from subjects who played

Table 1: Games in *Blues and Reds*.

| 2 rounds | 3 rounds | 4 rounds | 5 rounds | 6 rounds |
|---|---|---|---|---|
| 2.2 | 2.2.2 | 2.2.2.2 | 2.2.2.2.2 | 2.2.2.2.2.2 |
| 2.3 | 2.2.3 | 3.2.2.2 | 3.2.2.2.2 | |
| 2.4 | 2.3.2 | 4.2.2.2 | 4.2.2.2.2 | |
| 3.2 | 2.3.3 | 2.3.2.2 | | |
| 3.3 | 3.2.2 | 2.4.2.2 | | |
| | 3.2.3 | 2.2.3.2 | | |
| | 3.3.2 | 2.2.4.2 | | |
| | 3.3.3 | 2.2.2.3 | | |
| | 4.2.2 | 2.2.2.4 | | |

at least one game of *Blues and Reds* beyond the mandatory tutorial. For each game and each subject, we collected the following data.

1. At each round, response time (RT), which measures how many seconds a subject spends on selecting an action.

2. Whether a subject wins or loses.

**1. Response times.** Given the large size of our data, we take a conservative approach to remove outliers that can upward bias our metrics. For each tree game we keep observations in which the sum of RTs across rounds (henceforth Total Time, $TT$) is below the 95th percentile of its sample. Our final sample of tree games consists of 6,677 subjects who played 44,113 tree games. For each non-tree game, we keep observations in which Total Time is below the 90th percentile of its sample. In addition, in our analysis with non-tree games we only use data for subjects who won the equivalent tree game (this is part of our empirical strategy, which is explained in the final paragraphs of this section). Consequently, the data we use to analyze non-tree games is smaller than the one we use to analyze tree games. Our final sample of non-tree games consists of 4,582 subjects who played 26,997 non-tree games. Based on the observed RTs, we created the following variables that we use in our empirical analysis.

- $TT$. Total time a subject spends on solving an interaction. $TT$ is the sum of RTs from all rounds.

- $RT1$. Subject's response time at the first round of interaction. It is RT from the first round.
- $RRT1$. Subject's relative response time at the first round of interaction. It is $RT1$ as the percentage of $TT$; i.e., $RRT1 = \frac{RT1}{TT} \times 100\%$.
- $ATT$. Average total time spent on solving an interaction. Suppose that $N$ subjects participated in a given interaction. Let $TT_i$ be $TT$ of subject $i$. Then $ATT = \frac{1}{N}\sum_{i=1}^{N} TT_i$.
- $ART1$. Average response time at the first round of a given interaction. Suppose that $N$ subjects participated in a given interaction. Let $RT1_i$ be $RT1$ of subject $i$. Then $ART1 = \frac{1}{N}\sum_{i=1}^{N} RT1_i$.

**2. Subject wins.** We use tree games to study backward induction. Given the properties of tree games, we say that a subject who won in a tree was (behaving as if) backward inducting in that tree.

> ***Subject backward inducts.*** *We say that a subject behaves as if backward inducting in a tree game if she wins in that tree game.*

We use pairs of equivalent tree and non-tree games to study tree construction. Given the properties of non-tree games, winning in a non-tree means that a subject both constructed a tree and backward inducted. That is, losing in a non-tree means that the joint hypothesis of tree construction and backward induction is rejected; however, it does not allow to point at the cause.

Since our goal is to test the individual hypothesis of a subject constructing a tree, we deploy the same two-step approach as in Levitt et al. (2011). Their objective is to determine how backward inducting (BI) subjects behave in the centipede game. To that end, in the first step, they use the race game as a screening test to identify subjects who "demonstrated ability to backward induct flawlessly". In the second step, they look at the behavior of the selected subjects in the centipede game.

In our case, in the first step, we look at the behavior in the tree games that serves as our screening tests. For each tree, we select subjects who win; these are our BI subjects. In the second step, we turn to non-tree games. In a given non-tree game, we restrict our attention to BI subjects from the equivalent tree game. If a BI subject wins in a non-tree, then we say that she constructed a tree.

> ***Subject constructs a tree.*** *We say that a subject behaves as if constructing a tree in a non-tree game if she wins in that non-tree game and, previously, also won in the equivalent tree game.*

# 3 Do people construct trees and backward induct?

Our study begins with the obligatory analysis of whether subjects construct trees and backward induct. Although perception (i.e., tree construction) takes place before the selection of strategy (i.e., backward induction), our empirical analysis in this and the following two sections goes in the reverse order: first, we look at backward induction, then we study tree construction. This is because, as explained in section 2, the empirical strategy we deploy for the study of tree construction requires a prior analysis of backward induction.

## 3.1 Backward induction

As explained in section 2, a subject who loses in a tree game did not backward induct in that tree. For each of 27 trees, we provide the number of subjects who played the tree, $N$, and compute the failure rate of backward induction, that is, the percentage of subjects who did not backward induct, $\%NOT.BI$. This failure rate ranges from 2.5% to 53.1% (Table 2), while the theory predicts it should be zero.

Table 2 not only confirms what is already known in the literature (i.e., people do not backward induct) but also points at what causes the failures of backward induction. First, it must be the case that subjects differ in terms of their skills. If subjects were identical, then all subjects would perform equally well or equally poorly. That is, we would observe only two values of failure rate, 0% or 100%.

Second, observe that the percentage of subjects who violate backward induction is not the same across different trees. This indicates that the trees are not equally complex to our subjects; if they were, then the percentage of those who do not backward

Table 2: Percentage of subjects who did not backward induct.

| tree | $N$ | $\%NOT.BI$ |
|---:|:---:|:---:|
| 2.3 | 1,681 | 2.50% |
| 2.4 | 1,632 | 3.00% |
| 3.2 | 1,670 | 3.29% |
| 2.2 | 1,683 | 4.40% |
| 2.2.2 | 1,638 | 6.29% |
| 2.2.3 | 1,729 | 6.42% |
| 2.3.3 | 1,637 | 6.72% |
| 3.3 | 1,621 | 6.97% |
| 2.3.2 | 1,630 | 7.98% |
| 3.2.3 | 1,628 | 8.91% |
| 3.2.2 | 1,666 | 9.18% |
| 3.3.2 | 1,647 | 10.14% |
| 3.3.3 | 1,638 | 10.32% |
| 4.2.2 | 1,717 | 10.54% |
| 2.2.2.4 | 1,602 | 17.04% |
| 2.4.2.2 | 1,641 | 19.38% |
| 2.2.2.3 | 1,610 | 20.93% |
| 2.2.3.2 | 1,674 | 22.82% |
| 2.2.4.2 | 1,673 | 27.26% |
| 2.2.2.2.2 | 1,545 | 27.64% |
| 4.2.2.2 | 1,614 | 29.62% |
| 3.2.2.2 | 1,575 | 30.29% |
| 3.2.2.2.2 | 1,550 | 32.97% |
| 2.2.2.2 | 1,660 | 33.37% |
| 2.3.2.2 | 1,606 | 43.52% |
| 4.2.2.2.2 | 1,566 | 52.04% |
| 2.2.2.2.2.2 | 1,580 | 53.10% |

induct would be the same in each tree.

## 3.2 Tree construction

In terms of empirical strategy and tools that we use to understand tree construction, we follow the same steps as with the analysis of backward induction. The only aspect that changes is the data. As explained in section 2, we say that a subject who wins in a tree game but loses in the equivalent non-tree game does not construct a tree in that non-tree game. That is, to test for tree construction, we look at the behavior in non-tree games of subjects who backward inducted in the equivalent tree games.

For each of 27 non-trees, we provide the number of backward inducting subjects who played the non-tree, $N$, and compute the failure rate of tree construction, that is, the percentage of backward-inducting subjects who did not construct a tree, $\%NOT.TC$. This failure rate ranges from 12.25% to 63.95% (Table 3), while the theory predicts it should be zero.

As in the case of backward induction, we observe that, in general, subjects do not

Table 3: Percentage of backward-inducting subjects who did not construct a tree.

| non-tree | $N$ | $\%NOT.TC$ |
|---:|:---:|:---:|
| 2.2.3 | 1,184 | 12.25% |
| 2.2.2 | 1,149 | 14.01% |
| 3.3 | 1,265 | 14.39% |
| 2.3 | 1,348 | 18.03% |
| 2.4 | 1,273 | 18.70% |
| 2.3.2 | 1,114 | 19.57% |
| 2.3.3 | 1,214 | 19.60% |
| 2.2 | 1,343 | 20.70% |
| 3.2.3 | 1,104 | 23.55% |
| 3.2.2 | 1,133 | 34.77% |
| 3.3.2 | 1,119 | 35.39% |
| 3.2 | 1,344 | 35.94% |
| 4.2.2 | 1,112 | 36.15% |
| 2.3.2.2 | 721 | 36.34% |
| 3.3.3 | 1,115 | 37.13% |
| 2.4.2.2 | 1,041 | 37.18% |
| 2.2.2.3 | 967 | 41.88% |
| 2.2.3.2 | 957 | 45.35% |
| 2.2.2.2 | 834 | 45.44% |
| 2.2.2.4 | 1,012 | 46.64% |
| 2.2.4.2 | 886 | 49.77% |
| 2.2.2.2.2 | 644 | 50.47% |
| 3.2.2.2 | 816 | 51.96% |
| 4.2.2.2.2 | 405 | 52.35% |
| 4.2.2.2 | 861 | 57.49% |
| 2.2.2.2.2.2 | 434 | 58.29% |
| 3.2.2.2.2 | 602 | 63.95% |

construct trees. We also note that non-trees differ in complexity (because the failure rates differ) and subjects differ in their skills (because the failure rates are strictly between 0% and 100%). The variation in complexity and skills is what drives the failure of tree construction.

**Why to study tree construction?** Given how big the experimental literature on backward induction is, we do not find it necessary to motivate the importance of testing backward induction. However, since very little has been said about tree construction, we are compelled to emphasize the importance of testing whether people model the reality in a way that game theory postulates.

With no doubt, a tree is a very powerful tool to simplify and capture reality. With just dots and lines, a tree paints an image worth more than a thousand words, but it is only decision-makers in academic papers and textbooks who have the privilege of dealing with trees. Real-world interactions are depicted as anything but trees.

Creating models, not to mention constructing trees, is neither a trivial nor natural task. While game theory tells how to solve the trees (backward induction), game theory does not explain how to create the very same trees. Provided with the same tree, two economists would find the same equilibria; however, there is no reason to expect that they would translate the underlying interaction into the same tree. We should not assume that people analyze interactive situations as game theory does. This implies that neglecting the question of whether people perceive interactive situations in accordance with theory has serious consequences. In particular, if people do not construct trees, then testing backward induction could lead to erroneous conclusions.

# 4 Why don't people construct trees and backward induct?

In section 3, we learned that our subjects sometimes construct trees and sometimes backward instruct, but not always. We argued that the violations of game theory are driven by the subjects' skills (being too low) and the complexity of the games (being too high). In this section, we study skills and complexity.

## 4.1 Backward induction

### 4.1.1 Backward induction: skills

People differ in their abilities to backward induct. Higher skills imply a higher probability of a subject backward inducting. However, skills are not directly observable. What is observable is the choices reflecting those skills. Our objective is to figure out how to measure skills using the concept of response time (RT).

To elaborate on the problem of designing an RT-based measure of skills in dynamic games, consider the following fictional example. Take the 4-round tree 2.2.4.2 and four subjects. Their RTs at the first and third rounds are listed in Table 4.

According to the total time (TT), Ann and Chris are fast thinkers (each spent 20 seconds) while Bob and David are slow thinkers (each spent 40 seconds). If we look at the response time at the first round (RT1), Chris is the fastest thinker followed by Ann, David, and Bob. Finally, Ann and Bob allocate 75% of the total time at the first round while Chris and David allocate only 40%. How will the subjects in Table

Table 4: Response times of four fictional subjects in the tree 2.2.4.2.

| subject | $RT1$ | $RT3$ |
|---|---|---|
| Ann | 15 sec | 5 sec |
| Bob | 30 sec | 10 sec |
| Chris | 8 sec | 12 sec |
| David | 16 sec | 24 sec |

4 rank in terms of their likelihood to backward induct in 2.2.4.2?[13]

The key measure is the relative response time at the first round, denoted as $RRT1$ and computed as the time spent at the first round $RT1$ as the percentage of total time spent $TT$; i.e., $RRT1 = \frac{RT1}{TT} \times 100\%$. The higher $RRT1$, the more likely a subject is to backward induct. This is because when it comes to backward induction, what matters is the time spent at the first round. If a subject understands how to play — and this process of understanding takes place at the very beginning of a tree — then she does not have to analyze too much in the following rounds; she follows the strategy that she has already designed. This is especially true in our trees where making a mistake at any round is equivalent with losing a game.

In order to prove that the relative time spent at the first round is the key measure, we conduct the following empirical exercise. We consider three candidates for the measure of skills: $TT$ (total time), $RT1$ (response time at the first round), and $RRT1$ (relative response time at the first round).

For each tree, we divide the subjects into terciles by the corresponding measure. For example, take $TT$. The Low tercile corresponds to all subjects having $TT$ less than or equal to the percentile 33.3%. The Medium tercile corresponds to the subjects having $TT$ higher than the 33.3% percentile and less than or equal to the 66.6% percentile. Finally, the High $TT$ tercile corresponds to subjects with $TT$ higher than the 66.6%

[13] Intuition might indicate that it is the total time or just the response time at the first round (where analyzing a tree is most crucial) that is the best predictor of whether a subject behaves in accordance with game theory. After all, more time spent should be better. However, this intuition need not be entirely correct. Total time and time spent at the first round can be misleading indicators because spending more time on analyzing something a subject does not understand need not lead to a better outcome. For example, a professional game theorist would spend less time (both $TT$ and $RT1$) and is more likely to backward induct compared to someone who has never heard of game theory.

percentile.

For each tercile, we compute the percentage of backward-inducting subjects in that tercile. For instance, for the tree 2.2.2 and measure $RRT1$, 84% of subjects in the Low tercile backward inducted, 98% of subjects in the Medium tercile backward inducted, and 99% of subjects in the High tercile backward inducted.

The results are presented in Table 5. Since we are interested in relative time, our analysis excludes 2-round trees where $RRT1 = 100\%$ by default.

We observe that $RRT1$ is the best predictor of subjects backward inducting. This measure never fails in a sense that for each tree, higher $RRT1$ is associated with higher probability of backward inducting.

Table 5: Candidates for the measure of skills (backward induction).

| | $RRT1$ | | | $TT$ | | | $RT1$ | | |
|---|---|---|---|---|---|---|---|---|---|
| tree | L | M | H | L | M | H | L | M | H |
| 2.2.2 | 84% | 98% | 99% | 98% | 95% | 88% | 93% | 94% | 94% |
| 2.2.2.2 | 18% | 85% | 98% | 80% | 62% | 59% | 50% | 66% | 84% |
| 2.2.2.2.2 | 40% | 80% | 97% | 75% | 71% | 71% | 59% | 74% | 85% |
| 3.3.2 | 74% | 98% | 98% | 93% | 94% | 82% | 86% | 93% | 90% |
| 2.2.2.4 | 58% | 93% | 98% | 80% | 85% | 83% | 72% | 87% | 90% |
| 3.3.3 | 71% | 98% | 99% | 93% | 92% | 84% | 82% | 95% | 92% |
| 4.2.2.2 | 28% | 86% | 97% | 67% | 73% | 72% | 50% | 76% | 84% |
| 2.3.3 | 82% | 99% | 99% | 97% | 95% | 88% | 92% | 94% | 94% |
| 3.2.3 | 76% | 98% | 99% | 94% | 95% | 85% | 86% | 93% | 93% |
| 3.2.2.2 | 23% | 89% | 97% | 64% | 74% | 72% | 48% | 75% | 85% |
| 2.3.2 | 82% | 96% | 99% | 97% | 95% | 85% | 91% | 94% | 92% |
| 2.3.2.2 | 6% | 65% | 98% | 59% | 55% | 55% | 33% | 59% | 78% |
| 3.2.2 | 75% | 98% | 98% | 96% | 91% | 85% | 86% | 93% | 93% |
| 2.2.2.3 | 44% | 94% | 98% | 80% | 83% | 74% | 68% | 83% | 86% |
| 2.2.3 | 84% | 97% | 99% | 96% | 95% | 89% | 90% | 97% | 94% |
| 2.2.3.2 | 39% | 93% | 99% | 78% | 78% | 76% | 60% | 83% | 88% |
| 2.2.2.2.2.2 | 21% | 37% | 83% | 33% | 45% | 64% | 27% | 41% | 73% |
| 2.4.2.2 | 52% | 92% | 99% | 77% | 83% | 82% | 67% | 84% | 92% |
| 3.2.2.2.2 | 30% | 76% | 96% | 59% | 66% | 77% | 40% | 73% | 87% |
| 4.2.2 | 72% | 98% | 98% | 95% | 91% | 82% | 83% | 92% | 93% |
| 2.2.4.2 | 30% | 89% | 98% | 76% | 76% | 67% | 61% | 76% | 82% |
| 4.2.2.2.2 | 13% | 42% | 89% | 35% | 42% | 67% | 23% | 44% | 77% |

As of $TT$, in trees 2.2.2, 2.2.2.2, 2.2.2.2.2, 3.3.3, 2.3.3., 2.3.2, 2.3.2.2, 3.2.2, 2.2.3, 2.2.3.2, 4.2.2, and 2.2.4.2, higher $TT$ implies a subject being less likely to backward induct. For the trees 2.2.2.2.2.2, 3.2.2.2.2, and 4.2.2.2.2 we have the opposite relationship: higher $TT$ makes backward induction more likely. In addition, for the remaining seven trees, the relationship between $TT$ and the likelihood of backward induction is non-monotonic. In short, TT does not help us with ranking decision-makers in terms of their likelihood of backward inducting. The same is true for $RT1$.

According to $RRT1$, in our example in Table 4, Ann and Bob are equally good at backward induction and better than Chris and David, who share the same $RRT1$. Is there any way to refine our ranking and compare Ann vs Bob as well as Chris vs David? It turns out there is.

While $TT$ does not allow us to rank the subjects by itself alone, it does help once we condition on $RRT1$: for a given $RRT1$, higher $TT$ makes backward induction less likely. In other words, total time spent on solving a tree refines our key measure of skills. Consequently, in Table 4, Ann is better at backward induction than Bob, and Chris is better than David.

To analyze the impact of $TT$ conditional on $RRT1$, first, for a given tree, we divide the observations into terciles by $RRT1$. Then, we divide each $RRT1$-tercile into terciles with respect to $TT$. Therefore, for each tree we have 9 pairs $(TT_i|RRT1_j)$, where $i =$ High, Medium, Low and $j =$ High, Medium, Low. Table 6 below presents the results (percentage of subjects who backward inducted) for each of the nine groups across all trees in which the subject moves at least twice.

Table 6: Backward induction and $TT$ conditional on $RRT1$: aggregate analysis.

| | | $TT$ | | |
|---|---|---|---|---|
| | | L | M | H |
| $RRT1$ | L | 64.67% | 53.46% | 33.33% |
| | M | 92.53% | 89.71% | 78.15% |
| | H | 98.39% | 97.81% | 95.04% |

We observe that for each $RRT1$ tercile, increasing $TT$ decreases the proportion of subjects who backward induct. Importantly, Appendix Table 1 shows that this result also holds for individual trees with the following exceptions: Low, Medium, and High

$RRT1$-terciles in tree 2.2.2.4, and High $RRT1$-tercile in tree 2.4.2.2.

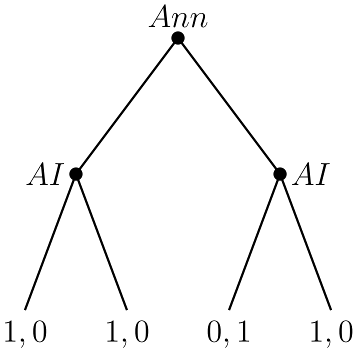
We believe that our result for $TT$ can be explained in a simple way. Consider Ann and Bob from Table 4. Both spend 75% of the total time at the first round, which would indicate that they are equally good at backward induction. However, Ann is faster (i.e., low $TT$) compared to Bob because she does not have to think too much about solving a tree. Bob is slower because it is not a typical problem for him as he is not familiar with the world of trees.[14]

To summarize, when it comes to the measure of skills in the context of backward induction, we find the following. The higher the relative time spent at the first round ($RRT1$), the more likely a subject is to backward induct. In addition, for a given value of $RRT1$, the higher the total time spent on solving the tree ($TT$), the less likely a subject is to backward induct.

### 4.1.2 Backward induction: complexity

As our results in section 3 indicate, trees vary in their complexity. In some cases, it is possible to say that one tree is objectively more complex than another tree. To elaborate on what "objective" means, consider the three trees depicted below, where Ann plays against AI and the payoffs are in the order $(Ann, AI)$.

Figure 1: Tree 2.2



[14] It is also possible that Bob is over-thinking (Gill and Prowse (2017)), but this is not a hypothesis we can test with our data.

Figure 2: Tree 3.3

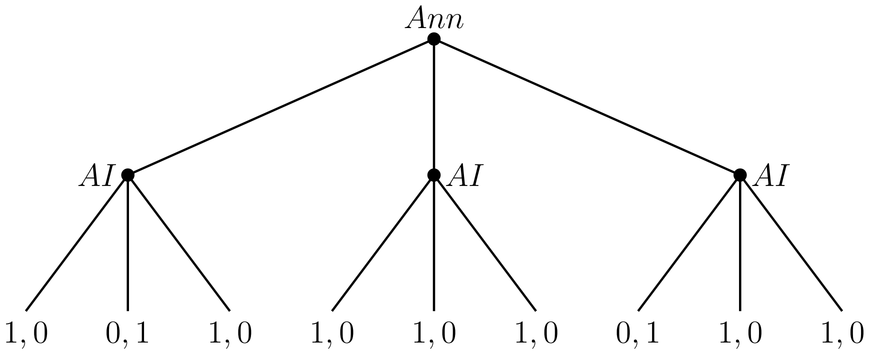
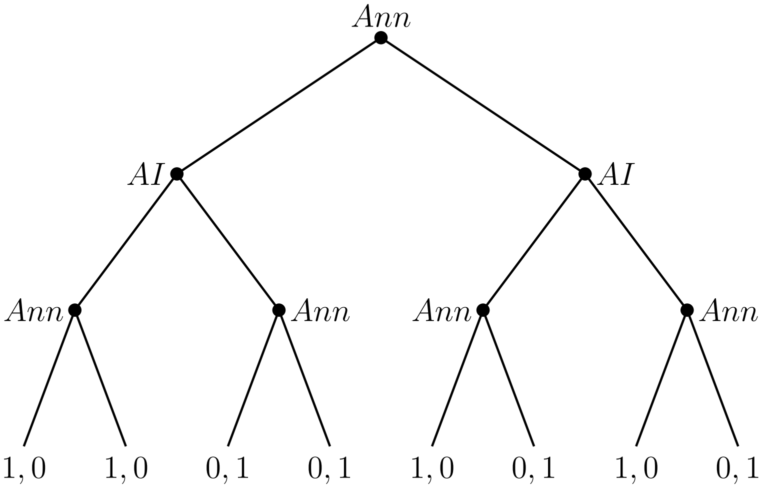


Figure 3: Tree 2.2.2



We can obtain tree 3.3 and tree 2.2.2 by adding nodes and branches to tree 2.2. That is, tree 2.2 is objectively less complex than both trees 3.3 and 2.2.2. More formally, we say that tree $N_1.N_2.N_3.N_4.N_5.N_6$ is objectively more complex than tree $M_1.M_2.M_3.M_4.M_5.M_6$ if

1. $N_i \geq M_i$ for each $i$ with at least one inequality being strict, or
2. tree $M_1.M_2.M_3.M_4.M_5.M_6$ consists of $k$ rounds (i.e., $M_l = 0$ for $l > k$) and there exists $j$ such that $N_j \geq M_1$, $N_{j+1} \geq M_2$ ,..., $N_{j+k} \geq M_k$ with at least one inequality being strict.

Out of 351 comparable pairs of trees, we have 136 pairs in which one tree is objectively more complex than another. All objective pairwise comparisons of complexity among the trees in our sample are depicted in Appendix Figure 8.

The main problem with objective complexity is that it does not generate a complete ranking of trees. For example, it is not clear whether tree 3.3 (Figure 2) is objectively more or less complex than tree 2.2.2 (Figure 3). For this reason, we need an empirical measure of complexity. What we are looking for is a measure that satisfies three conditions: (a) it is complete (i.e., ranks any pair of trees), (b) it is the best extension of the objective measure (i.e., agrees with the objective measure for as many pairwise comparisons as possible), and (c) it is based on data (i.e., empirical measure).

In Table 7, we rank the trees in accordance with the three potential empirical measures of complexity: $\%NOT.BI$ (percentage of subjects who did not backward induct), $ATT$ (average total time), and $ART1$ (average response time at the first round).

In order to identify the best extension of the objective measure, we conduct the following exercise. If tree A is objectively more complex than tree B, then we check whether a candidate for the empirical measure of complexity ranks tree A higher than tree 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 objective measure. If tree A is objectively more complex than tree 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 objective measure disagree. Finally, if tree A is objectively more complex than tree 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. The results are depicted in Table 8.

Looking at Table 8, we conclude that $ATT$ and $ART1$ are very similar and far superior to $\%NOT.BI$. In fact, $ATT$ and $ART1$ almost perfectly replicate the objective ranking of complexity. Consequently, we consider both $ATT$ and $ART1$ as the appropriate empirical measures of complexity of backward induction. In the rest of our paper, we present the results only for $ART1$ as our empirical measure of complexity; qualitative results with $ATT$ as the empirical measure of complexity are the same.

To complement Table 8, we present the heat map in Appendix Figure 9, which cap-

Table 7: Candidates for the empirical measure of complexity (backward induction).

| tree | $\%NOT.BI$ | tree | $ATT$ | tree | $ART1$ |
|---|---|---|---|---|---|
| 2.3 | 2.50% | 2.2 | 9.09 | 2.2 | 9.09 |
| 2.4 | 3.00% | 2.3 | 9.47 | 2.3 | 9.47 |
| 3.2 | 3.29% | 2.4 | 9.53 | 2.4 | 9.53 |
| 2.2 | 4.40% | 3.2 | 9.74 | 3.2 | 9.74 |
| 2.2.2 | 6.29% | 3.3 | 10.51 | 3.3 | 10.51 |
| 2.2.3 | 6.42% | 2.2.2 | 19.05 | 2.2.2 | 13.53 |
| 2.3.3 | 6.72% | 3.2.2 | 21.03 | 2.3.2 | 15.03 |
| 3.3 | 6.97% | 2.3.2 | 21.04 | 2.2.3 | 15.70 |
| 2.3.2 | 7.98% | 2.2.3 | 21.07 | 3.2.2 | 15.75 |
| 3.2.3 | 8.91% | 3.2.3 | 21.94 | 3.2.3 | 16.92 |
| 3.2.2 | 9.18% | 4.2.2 | 22.68 | 4.2.2 | 17.31 |
| 3.3.2 | 10.14% | 2.3.3 | 22.98 | 3.3.2 | 17.91 |
| 3.3.3 | 10.32% | 3.3.2 | 23.18 | 2.3.3 | 18.11 |
| 4.2.2 | 10.54% | 3.3.3 | 25.64 | 3.3.3 | 20.53 |
| 2.2.2.4 | 17.04% | 2.2.2.2 | 30.10 | 2.2.2.2 | 21.61 |
| 2.4.2.2 | 19.38% | 2.2.2.4 | 32.57 | 2.3.2.2 | 26.23 |
| 2.2.2.3 | 20.93% | 2.2.2.3 | 33.24 | 2.2.2.3 | 26.26 |
| 2.2.3.2 | 22.82% | 2.2.3.2 | 34.55 | 2.2.2.4 | 26.37 |
| 2.2.4.2 | 27.26% | 2.4.2.2 | 35.84 | 2.2.3.2 | 26.99 |
| 2.2.2.2.2 | 27.64% | 2.3.2.2 | 35.98 | 2.4.2.2 | 28.89 |
| 4.2.2.2 | 29.62% | 3.2.2.2 | 38.44 | 3.2.2.2 | 31.23 |
| 3.2.2.2 | 30.29% | 4.2.2.2 | 38.63 | 4.2.2.2 | 31.26 |
| 3.2.2.2.2 | 32.97% | 2.2.4.2 | 40.44 | 2.2.4.2 | 31.59 |
| 2.2.2.2 | 33.37% | 2.2.2.2.2 | 59.43 | 2.2.2.2.2 | 43.33 |
| 2.3.2.2 | 43.52% | 3.2.2.2.2 | 66.26 | 3.2.2.2.2 | 48.92 |
| 4.2.2.2.2 | 52.04% | 2.2.2.2.2.2 | 81.55 | 2.2.2.2.2.2 | 56.16 |
| 2.2.2.2.2.2 | 53.10% | 4.2.2.2.2 | 95.08 | 4.2.2.2.2 | 72.27 |

tures how our empirical measure of complexity extends the objective measure of complexity from Appendix Figure 8. In the heat map, trees are ranked according to their value of $ART1$ from the least complex (2.2) to the most complex (4.2.2.2.2). Colors indicate statistical significance of the difference in empirical complexities. Red, orange, and yellow mean that the $ART1$ of the game in the Y-axis is higher and statistically different than that of the game in the X-axis at 1%, 5%, and 10%, re-

Table 8: Selecting the empirical measure of complexity (backward induction).

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

spectively. Gray means that there is no statistical difference between $ART1$s of two games. For example, the row corresponding to game 4.2.2.2 indicates that its $ART1$ is statistically different at the 1% level with respect to all games with a lower value of $ART1$ except for 3.2.2.2, where the difference is not significant.

Our next objective is to analyze the main drivers of empirical complexity in the case of backward induction. In particular, we look for connections between the tree structure and empirical complexity.

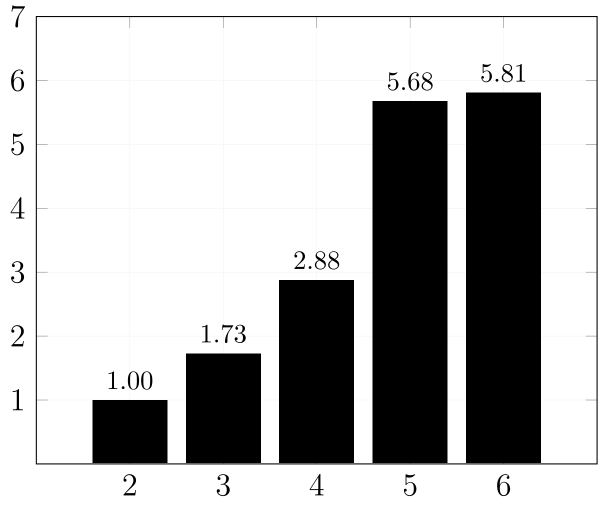
Starting with a given tree, we can add more rounds (make the tree longer) or add more actions at the already existing nodes (make the tree wider). The three questions we ask are the following:

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. Table 1 presents the trees grouped by the number of rounds. Figure 4 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 4 clearly shows that increasing the number of rounds increases the empirical complexity of a tree.[15]

[15] 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

Figure 4: Empirical complexity and length (backward induction).



We now study the impact of increasing the number of actions per round. The analysis is presented in Table 9.

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 9 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.[16]

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.

[16] 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.

Table 9: Empirical complexity and width (backward induction).

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<br>3.2 | 2.3.2<br>3.2.2<br>2.2.3 | 3.2.2.2<br>2.3.2.2<br>2.2.3.2<br>2.2.2.3 | 3.2.2.2.2 |
| Max | 3.3<br>2.4 | 3.3.2<br>2.3.3<br>3.2.3<br>4.2.2<br>3.3.3 | 2.2.2.4<br>4.2.2.2<br>2.4.2.2<br>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 |

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 Appendix Figure 10, we graphically depict our analysis and discuss it in detail below.

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.

## 4.2 Tree construction

### 4.2.1 Tree construction: skills

As is the case of backward induction, people also differ in their abilities to construct trees. Again, our challenge is to design the measure of skills. We proceed as we did in the case of analyzing backward induction. In Table 10, we replicate the analysis we conducted for backward induction in Table 5. We observe that, as was the case in our analysis of backward induction, the best measure of skills in the context of tree construction is the relative response time at the first round ($RRT1$): the percentage of subjects constructing trees increases from the low to the high tercile across all non-trees. However, for $TT$ and $RT1$, the relationship between the percentage of subjects constructing trees and a measure is not the same for each non-tree.

Next, we turn to analyze the conditional behavior of total time ($TT$). In Table 11, we replicate Table 6 for tree construction. We first divide all the data into terciles by $RRT1$. Then, we divide each tercile into additional terciles with respect to $TT$.

We observe that for each $RRT1$ tercile, increasing $TT$ decreases the proportion of subjects who construct a tree. In Appendix Table 2, we show that this holds in most non-trees individually, although the result is not as sharp as in the case of games represented as trees.

To summarize, when it comes to the measure of skills in the context of tree construction, we find the following. The higher the relative time spent at the first round ($RRT1$), the more likely a subject is to construct a tree. In addition, for a given value of $RRT1$, the higher the total time spent on solving the non-tree ($TT$), the less likely a subject is to construct a tree.

### 4.2.2 Tree construction: complexity

In order to design the empirical measure of complexity in the context of tree construction, we follow the same strategy as we did for the case of analyzing backward induction.

As with backward induction, we also want our measure of complexity to be the best

Table 10: Candidates for the measure of skills (tree construction).

| | $RRT1$ | | | $TT$ | | | $RT1$ | | |
|---|---|---|---|---|---|---|---|---|---|
| non-tree | L | M | H | L | M | H | L | M | H |
| 2.2.2 | 72% | 89% | 97% | 94% | 87% | 77% | 84% | 89% | 85% |
| 2.2.2.2 | 17% | 56% | 91% | 41% | 62% | 61% | 29% | 65% | 70% |
| 2.2.2.2.2 | 25% | 43% | 81% | 42% | 51% | 56% | 34% | 47% | 68% |
| 3.3.2 | 37% | 64% | 93% | 76% | 59% | 58% | 56% | 63% | 75% |
| 2.2.2.4 | 23% | 44% | 93% | 38% | 57% | 65% | 26% | 58% | 76% |
| 3.3.3 | 19% | 74% | 96% | 67% | 62% | 59% | 43% | 68% | 77% |
| 4.2.2.2 | 12% | 31% | 85% | 26% | 37% | 64% | 13% | 44% | 71% |
| 2.3.3 | 55% | 88% | 98% | 91% | 83% | 68% | 75% | 82% | 84% |
| 3.2.3 | 53% | 85% | 92% | 83% | 83% | 63% | 72% | 82% | 75% |
| 3.2.2.2 | 15% | 40% | 89% | 33% | 49% | 62% | 24% | 48% | 73% |
| 2.3.2 | 58% | 86% | 96% | 85% | 82% | 74% | 72% | 84% | 86% |
| 2.3.2.2 | 22% | 74% | 95% | 46% | 62% | 84% | 30% | 71% | 90% |
| 3.2.2 | 22% | 79% | 95% | 79% | 65% | 51% | 54% | 69% | 73% |
| 2.2.2.3 | 24% | 58% | 93% | 44% | 63% | 68% | 33% | 63% | 77% |
| 2.2.3 | 74% | 93% | 97% | 93% | 89% | 81% | 84% | 89% | 90% |
| 2.2.3.2 | 17% | 53% | 94% | 50% | 54% | 60% | 26% | 64% | 74% |
| 2.2.2.2.2.2 | 17% | 26% | 83% | 16% | 49% | 60% | 16% | 37% | 72% |
| 2.4.2.2 | 34% | 59% | 96% | 55% | 61% | 72% | 40% | 66% | 83% |
| 3.2.2.2.2 | 11% | 23% | 74% | 25% | 32% | 51% | 15% | 34% | 60% |
| 4.2.2 | 22% | 75% | 94% | 69% | 61% | 61% | 47% | 65% | 79% |
| 2.2.4.2 | 11% | 50% | 90% | 29% | 56% | 67% | 18% | 56% | 76% |
| 4.2.2.2.2 | 21% | 36% | 87% | 38% | 44% | 61% | 27% | 41% | 75% |

Table 11: Tree construction and $TT$ conditional on $RRT1$: aggregate analysis.

| | | $TT$ | | |
|---|---|---|---|---|
| | | L | M | H |
| $RRT1$ | L | 45.43% | 30.59% | 22.72% |
| | M | 73.15% | 67.85% | 52.40% |
| | H | 96.23% | 92.80% | 88.55% |

extension of the objective measure. The ranking of non-trees according to their objective complexity is the same as the ranking of objective complexity of the equivalent trees (see Appendix Figure 8). Following the same approach as in the study of backward induction, we start with Table 12, in which we rank the non-trees in accordance with the three potential empirical measures of complexity: $\%NOT.TC$ (percentage of backward-inducting subjects who did not construct a non-tree), $ATT$ (average total time), and $ART1$ (average response time at the first round).

Table 12: Candidates for the empirical measure of complexity (tree construction).

| non-tree | $\%NOT.TC$ | non-tree | $ATT$ | non-tree | $ART1$ |
|---:|:---:|---:|:---:|---:|:---:|
| 2.2.3 | 12.25% | 2.3 | 12.04 | 2.3 | 12.04 |
| 2.2.2 | 14.01% | 2.2 | 12.69 | 2.2 | 12.69 |
| 3.3 | 14.39% | 2.4 | 12.77 | 2.4 | 12.77 |
| 2.3 | 18.03% | 3.3 | 14.05 | 3.3 | 14.05 |
| 2.4 | 18.70% | 3.2 | 18.02 | 3.2 | 18.02 |
| 2.3.2 | 19.57% | 2.2.3 | 29.96 | 2.2.3 | 21.54 |
| 2.3.3 | 19.60% | 2.2.2 | 30.29 | 2.2.2 | 22.03 |
| 2.2 | 20.70% | 2.3.3 | 39.01 | 2.3.3 | 27.49 |
| 3.2.3 | 23.55% | 2.3.2 | 40.64 | 2.3.2 | 30.24 |
| 3.2.2 | 34.77% | 3.2.3 | 41.65 | 2.2.2.3 | 30.76 |
| 3.3.2 | 35.39% | 2.2.2.3 | 42.86 | 3.2.2 | 31.10 |
| 3.2 | 35.94% | 2.2.2.2 | 43.82 | 3.2.3 | 31.35 |
| 4.2.2 | 36.15% | 3.2.2 | 43.95 | 2.2.2.2 | 32.53 |
| 2.3.2.2 | 36.34% | 3.3.2 | 45.57 | 3.3.2 | 32.71 |
| 3.3.3 | 37.13% | 4.2.2 | 45.62 | 4.2.2 | 33.69 |
| 2.4.2.2 | 37.18% | 3.3.3 | 49.79 | 3.3.3 | 36.46 |
| 2.2.2.3 | 41.88% | 2.2.2.4 | 50.23 | 2.2.2.4 | 36.49 |
| 2.2.3.2 | 45.35% | 2.2.4.2 | 54.58 | 2.2.3.2 | 38.25 |
| 2.2.2.2 | 45.44% | 3.2.2.2 | 55.31 | 2.4.2.2 | 40.37 |
| 2.2.2.4 | 46.64% | 2.2.3.2 | 55.47 | 2.2.4.2 | 40.73 |
| 2.2.4.2 | 49.77% | 2.3.2.2 | 57.11 | 3.2.2.2 | 41.82 |
| 2.2.2.2.2 | 50.47% | 2.4.2.2 | 58.09 | 2.3.2.2 | 43.12 |
| 3.2.2.2 | 51.96% | 4.2.2.2 | 60.51 | 4.2.2.2 | 45.54 |
| 4.2.2.2.2 | 52.35% | 2.2.2.2.2 | 96.84 | 2.2.2.2.2 | 58.58 |
| 4.2.2.2 | 57.49% | 2.2.2.2.2.2 | 107.87 | 2.2.2.2.2.2 | 66.92 |
| 2.2.2.2.2.2 | 58.29% | 3.2.2.2.2 | 109.63 | 3.2.2.2.2 | 71.17 |
| 3.2.2.2.2 | 63.95% | 4.2.2.2.2 | 162.24 | 4.2.2.2.2 | 116.82 |

Next, in order to pick the best measure of complexity, in Table 13, we replicate the analysis from Table 8.

Table 13: Selecting the empirical measure of complexity (tree construction).

| | Agree | Disagree | Undefined |
|---|---|---|---|
| $\%NOT.TC$ | 87 | 13 | 36 |
| $ATT$ | 125 | 3 | 8 |
| $ART1$ | 125 | 3 | 8 |

Table 13 tells us that $ATT$ and $ART1$ are the best measures of complexity. Hereafter, all the results are computed for $ART1$, but qualitative results remain the same if we use $ATT$ as our measure of complexity. To complement Table 13, we present the heat map in Appendix Figure 11 which captures how our empirical measure of complexity extends the objective measure of complexity from Appendix Figure 8.

With the measure of complexity in hand, we are ready to analyze the impact of the structure of interaction on its complexity. As with the analysis of backward induction, we can think of three questions we want to answer.

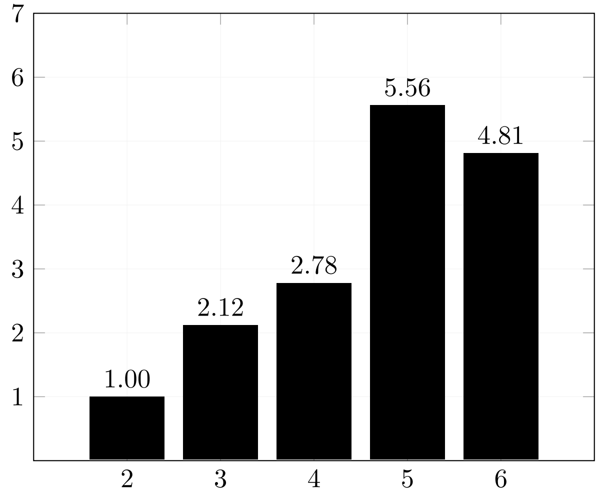
1. Is a longer non-tree more complex?
2. Is a wider non-tree more complex?
3. What increases the complexity of a non-tree more: length or width?

First, we analyze the impact of length. Figure 5, which replicates Figure 4, depicts the average complexity by the number of rounds and shows that increasing the number of rounds increases the empirical complexity of a non-tree.[17]

Second, we analyze the impact of width. Table 14 replicates the analysis we conducted in Table 9. We observe that with exception of the 2-round interactions (where

[17] Note that there is a decrease in complexity from the group of 5-round games to the 6-round game. 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, where the latter is by far the most complex game in our sample. 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 14%, from 58.58 to 66.92. In addition, we could also remove the most complex non-tree (viz. 4.2.2.2.2). This is because its individual complexity is 116.82 (i.e., on average, subjects spend 116.82 second on the first round of the non-tree 4.2.2.2.2) while the complexity of the second most complex non-tree (3.2.2.2.2) is 71.17. In this case, the index for 5-round games drops to 4.64, which is lower than the index for our 6-round game (4.81).

Figure 5: Empirical complexity and length (tree construction).



the complexity initially increases to decrease for the widest interactions), wider interactions are more difficult for the subjects to depict them as trees.

Finally, we analyze what has a bigger impact on making a non-tree interaction more complex: making it longer or making it wider. To that end, we conduct the same exercise we did for the same analysis in the context of backward induction. Appendix Figure 12 graphically depicts our analysis which we discuss in detail below.

Case 1. We start with the non-tree 2.2 (4 final nodes), whose empirical complexity is 12.69. We can make 2.2 longer by expanding it to 2.2.2 (8 final nodes); with this, the empirical complexity increases to 22.03. Alternatively, we can make 2.2 wider by expanding it to 2.4 (8 final nodes); the empirical complexity increases to 12.77. We observe that elongation is more important.

Case 2. We start with the non-tree 2.2.2 (8 final nodes), whose empirical complexity is 22.03. We can make 2.2.2 longer by expanding it to 2.2.2.2 (16 final nodes) with the empirical complexity 32.53. Alternatively, we can make 2.2.2 wider by expanding it to 4.2.2 (16 final nodes) with the empirical complexity 33.69. We observe that making an interaction wider seems to make the game relatively more complex. However, the difference in $ART1$ between 2.2.2.2 and 4.2.2 is not statistically significant (p-value = 0.19). In this case the results are not conclusive.

Case 3. We start with the non-tree 2.2.2.2 (16 final nodes), whose empirical com-

Table 14: Empirical complexity and width (tree construction).

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<br>3.2 | 2.3.2<br>3.2.2<br>2.2.3 | 3.2.2.2<br>2.3.2.2<br>2.2.3.2<br>2.2.2.3 | 3.2.2.2.2 |
| Max | 3.3<br>2.4 | 3.3.2<br>2.3.3<br>3.2.3<br>4.2.2<br>3.3.3 | 2.2.2.4<br>4.2.2.2<br>2.4.2.2<br>2.2.4.2 | 4.2.2.2.2 |

Panel (b)

| | 2 rounds | 3 rounds | 4 rounds | 5 rounds |
|---|---|---|---|---|
| Min | 12.69 | 22.03 | 32.53 | 58.58 |
| Med | 15.03 | 27.52 | 38.01 | 71.17 |
| Max | 13.41 | 32.26 | 40.59 | 116.82 |

plexity is 32.53. We can make 2.2.2.2 longer by expanding it to 2.2.2.2.2 (32 final nodes) with the empirical complexity 58.58. Alternatively, we can make 2.2.2.2 wider by expanding it to one of the following non-trees: 2.4.2.2, 2.2.4.2, 2.2.2.4, or 4.2.2.2. Each of these non-trees has 32 final nodes. Their average complexity is 40.78. We observe that elongation is more important.

Case 4. We start with the non-tree 2.2.2.2.2 (32 final nodes), whose empirical complexity is 58.58. 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 66.92. 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 116.82. We observe that elongation is less important. However, as already suggested in footnote 17, the non-tree 4.2.2.2.2 is an outlier: with the complexity of 116.82 it is the most complex non-tree, while the second most complex non-tree (3.2.2.2.2) has the complexity of 71.17.[18]

To summarize, we answer the three questions we asked as follows.

1. Is a longer non-tree more complex? Yes.
2. Is a wider non-tree more complex? Yes.
3. What increases the complexity of a non-tree more: length or width? Length seems more important than width, but results are not as conclusive as in the case of backward induction.

# 5 Complexity versus skills: what matters more?

## 5.1 Backward induction: complexity vs skills

We analyze the impact of skills and complexity on the probability of observing a subject backward inducting. To that end, we first run a logit regression to estimate $p_i = Pr(Y_i = 1|x_i) = \Lambda(x_i'\beta)$, where $Y_i$ is a binary variable capturing whether the subject $i$ backwards inducts or not; $x_i$ contains measures of a subject's skills, complexity of tree, and a control for the order in which the game appeared to the subject; and $\Lambda(s)$ is the logistic function $\frac{e^s}{1+e^s}$ . More precisely, we use Maximum Likelihood Estimation (MLE) to estimate the parameters of the following standard logit model.

$$Logit(Y) = \alpha + \beta X \tag{1}$$

$Y$ is the dependent variable in the regression capturing whether the subject backward inducted ($Y_i = 1$) or did not backward induct ($Y_i = 0$), $\alpha$ is the intercept, and $X$

[18] The large jump in complexity from 3.2.2.2.2 to 4.2.2.2.2 points towards the width becoming increasingly more relevant as length increases. As such, there might be a non-linear relationship between length and width as the number of rounds increases. In order to answer this and other important questions, we are currently developing our second mobile experiment.

is the $N \times 3$ matrix where $N$ is the number of observations in the sample and the three independent variables are Skills ($RRT1$), Complexity ($ART1$), and Sequence ($Seq$). The independent variable Sequence is the order in which a tree appeared in a subject's sequence of trees ($Seq = 1, ..., 27$). Table 15 depicts summary statistics of the independent variables.

Table 15: Summary statistics for the logit regression (backward induction).

| Variable | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|
| $Y$ | 0.78 | 0.41 | 0 | 1 |
| Skills | 0.74 | 0.16 | 0.02 | 0.98 |
| Complexity | 27.98 | 14.42 | 13.53 | 72.27 |
| Sequence | 8.20 | 7.22 | 1 | 27 |
| Number of observations = 35,826 | | | | |

Table 16 shows the results from the logit regression. Although the economic magnitude of the coefficients is hard to interpret, some important patterns arise. First, all coefficients are statistically significant. Second, the signs of the coefficients are what we expect: higher skills increase and higher complexity reduces the probability of a subject backward inducting. In addition, more experience (variable Sequence) also increases that probability.

We now asses how good this simple model is to predict whether or not a subject backwards inducts. For this purpose, we calculate 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 $\hat{p}_i < 0.5$. The model shows good predictive power: it correctly predicts whether or not a subject is going to backward induct in 86.55% of the cases. Moreover, the high rate of success comes from correctly predicting when a subject backward inducts, $Pr(Y = 1|\hat{Y} = 1) = 88.41\%$, as well as when she does not, $Pr(Y = 0|\hat{Y} = 0) = 76.25\%$.

We now move to analyze the economic magnitude of the different explanatory variables. To that end, we calculate the marginal effect of these variables (evaluated at their mean values). More precisely, for the case of the logit regression, the marginal effect is calculated as $\frac{\partial p}{\partial x_j} = \Lambda(\bar{x}'\hat{\beta})\{1 - \Lambda(\bar{x}'\hat{\beta})\}\hat{\beta}_j$, where $\bar{x}$ is the vector of mean independent variables, $\hat{\beta}$ is the vector of estimated coefficients, and $\hat{\beta}_j$ for

Table 16: Results from the logit regression (backward induction).

| Variable | Logit MLE |
|---|---|
| Skills | 9.945<br>(0.138) |
| Complexity | -0.037<br>(0.001) |
| Sequence | 0.042<br>(0.003) |
| Intercept | -4.939<br>(0.100) |
| Observations | 35,826 |
| Correct Prediction (%): Overall | 86.55% |
| Correct Prediction (%): Backward inducting | 88.41% |
| Correct Prediction (%): Not backward inducting | 76.25% |

Robust standard errors are in parenthesis. The data contains only tree games in which the subject has to move in two or more rounds.

$j = \{Skills, Complexity, Sequence\}$ is the estimated parameter of the variable for which the marginal effect is being calculated. In addition, to simplify the economic interpretation, we multiply the marginal effects of Skills and Complexity by their standard deviations. Results are in Table 17.

Table 17: Marginal effect from the logit regression (backward induction).

| Variable | Margins Estimate |
|---|---|
| Skills* | 20.04% |
| Complexity* | -6.60% |
| Sequence | 0.52% |

* Marginal effect calculated in terms of one-standard deviation.

We interpret the results in Table 17 in the following way. Increasing a subject's skills by one standard deviation increases the probability of backward induction by 20.04%. Decreasing the complexity of a tree by one standard deviation increases the probability of backward induction by 6.60%. Finally, if the tree appears in sequence $Seq$ instead of $Seq - 1$, the probability of backward induction increases by 0.52%. The comparison of relative importance between skills and complexity indicates that it is the skills that have a bigger impact on the probability of a subject backward inducting.

## 5.2 Tree construction: complexity vs skills

We analyze the relative impact of skills and complexity on a subject creating a tree. We use the same approach as in the case of backward induction. As a reminder, we use Maximum Likelihood Estimation (MLE) to estimate the parameters of the following standard logit model.

$$Logit(Y) = \alpha + \beta X \tag{2}$$

$Y$ is the dependent variable in the regression capturing whether the subject created a tree ($Y_i = 1$) or did not create a tree ($Y_i = 0$), $\alpha$ is the intercept, and $X$ is the $N \times 3$ matrix where $N$ is the number of observations in the sample and the three independent variables are Skills ($RRT1$), Complexity ($ART1$), and Sequence ($Seq$). The independent variable Sequence is the order in which a non-tree appeared in a subject's sequence of trees ($Seq = 1, ..., 27$). Table 18 depicts summary statistics of the independent variables.

Table 18: Summary statistics for the logit regression (tree construction).

| Variable | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|
| Win | 0.63 | 0.48 | 0.00 | 1.00 |
| Skills | 0.69 | 0.19 | 0.02 | 0.98 |
| Complexity | 37.81 | 15.58 | 21.54 | 116.82 |
| Sequence | 9.35 | 7.47 | 1 | 27 |
| Number of observations = 20,424 | | | | |

Table 19 shows the results from the logit regression: (1) all coefficients are statistically significant, (2) the signs of the coefficients are what we expect, and (3) using the cutoff value of 50%, the model shows good predictive power: it correctly predicts whether or not a subject is going to construct a tree in 78.30% of the cases. The high rate of success comes from correctly predicting when a subject creates a tree, $Pr(Y = 1|\hat{Y} =$

1) = 79.72%, as well as when she does not, $Pr(Y = 0|\hat{Y} = 0) = 75.00\%$.

Table 19: Results from the logit regression (tree construction).

| Variable | Logit MLE |
|---|---|
| Skills | 7.412<br>(0.125) |
| Complexity | -0.022<br>(0.001) |
| Sequence | 0.040<br>(0.003) |
| Intercept | -3.976<br>(0.095) |
| Observations | 20,424 |
| Correct Prediction (%): Overall | 78.30% |
| Correct Prediction (%): Constructing a tree | 79.72% |
| Correct Prediction (%): Not constructing a tree | 75.00% |

Robust standard errors are in parenthesis. The data contains only non-tree games in which the subject has to move in two or more rounds.

We now move to analyze the economic magnitude of the different explanatory variables. To that end, we calculate the marginal effects of these variables (evaluated at their mean values). In addition, to simplify the economic interpretation, we multiply the marginal effect of Skills and Complexity by their standard deviations. Results are in Table 20.

Table 20: Marginal effect from the logit regression (tree construction).

| Variable | Margins Estimate |
|---|---|
| Skills* | 32.18% |
| Complexity* | -7.84% |
| Sequence | 0.91% |

[*] Marginal effect calculated in terms of one-standard deviation.

Increasing a subject's skills by one standard deviation increases the probability of tree construction by 32.18%. Decreasing the complexity of an interaction by one standard deviation increases that probability by 7.84%. Finally, if a non-tree appears in sequence $Seq$ instead of $Seq - 1$, the probability of tree construction increases by 0.91%. The comparison of relative importance between skills and complexity indicates that it is the skills that have a bigger impact on a subject constructing a tree; in fact, that impact is bigger than it is in the case of backward induction.

## 6 Conclusions

In this paper, we study two fundamental concepts from game theory: how the subjects perceive an interaction they participate in (tree construction) and the algorithm they use to select strategies (backward induction). We find that sometimes people behave as predicted by theory (i.e., they construct trees and backward induct) and sometimes they do not.

We determine that the subject's skills and the complexity of the interaction are the key driving forces behind the violations of the theoretical predictions. Then, we propose new measures of a subject's skills and complexity of interaction; both are based on response times.

When it comes to skills, for both tree construction and backward induction, we find that how a subject allocates time when thinking is more important than the total thinking time she spends. More precisely, it is the relative response time at the first round ($RRT1$) that is the key measure of skills. The higher $RRT1$, the more likely a subject is to construct a tree and to backward induct. In addition, conditional on $RRT1$, the higher the total time ($TT$), the less likely a subject is to create a tree and to backward induct.

When it comes to complexity, we find that for both trees (which we use to analyze backward induction) and non-trees (which we use to analyze tree construction), the average total time ($ATT$) and the average response time at the first round ($ART1$) are the best measures of complexity. In the case of backward induction, making a tree longer or wider makes it more complex; we also find that the length is relatively more important. In the case of tree construction, making an interaction longer makes it more complex. The evidence that making an interaction wider makes it more complex in the case of tree construction is weaker than in the case of backward induction but still present. Additionally, in the context of complexity of tree construction, length seems more important than width, but results are not as conclusive as in the case of backward induction.

Finally, we compare the relative importance of skills and complexity on tree construction and on backward induction. In both cases, we find that the skills have a bigger impact on the likelihood of observing a subject backward inducting or constructing a tree. Consequently, improving skills is of primary importance, while simplification of complexity seems to be of second-order importance.

Our data comes from a mobile experiment. We created a mobile game, *Blues and Reds*, that has been globally available since August 2017 on iOS and Android devices. Our subjects come from over 100 countries. In the sample that we use to study backward induction, there are 6,677 subjects who played 44,113 tree games. In the sample that we use to study tree construction, there are 4,582 subjects who played 26,997 non-tree games.

# 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.

Arieli, A., Y. Ben-Ami, and A. Rubinstein (2011): "Tracking Decision Makers under Uncertainty," *American Economic Journal: Microeconomics*, 3, 68–76.

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 (2016a): "Logical abilities and behavior in strategic-form games," *Journal of Economic Psychology*, 56, 39–59.

——— (2016b): "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.

Brocas, I., J. D. Carrillo, S. W. Wang, and C. F. Camerer (2014): "Imperfect Choice or Imperfect Attention? Understanding Strategic Thinking in Private Information Games," *Review of Economic Studies*, 81, 944–970.

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 ob 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.

Cason, T. N. and C. R. Plott (2014): "Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing," *Journal of Political Economy*, 122, 1235–1270.

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.

Cox, J. C. and D. James (2012): "Clocks and Trees: Isomorphic Dutch Auctions and Centipede Games," *Econometrica*, 80, 883–903.

Crawford, V. P., M. A. Costa-Gomes, and N. Iriberri (2013): "Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications," *Journal of Economic Literature*, 51, 1–15.

Devetag, G., S. D. Guida, and L. Polonio (2016): "An Eye-tracking Study of Feature-based Choice in One-shot Games," *Experimental Economics*, 19, 177–201.

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.

Friedenberg, A., W. Kets, and T. Kneeland (2017): “Bounded Reasoning: Rationality or Cognition,” *working paper.*

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.*

Halevy, N., E. Chou, and J. K. Murnighan (2012): “Mind Games: The Mental Representation of Conflict,” *Journal of Personality and Social Psychology*, 102, 132–148.

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.

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.

Kneeland, T. (2015): "Identifying Higher-Order Rationality," *Econometrica*, 83, 2065–2079.

Knoepfle, D. T., C. F. Camerer, and J. Wang (2009): "Studying Learning in Games Using Eye-Tracking," *Journal of the European Economic Association*, 7, 388–398.

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

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.

McCabe, K. A., V. L. Smith, and M. LePore (2000): "Intentionality Detection and "Mindreading": Why Does Game Form Matter?" *Proceedings of the National Academy of Sciences*, 97, 4404–4409.

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.

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

Rapoport, A. (1997): "Order of Play in Strategically Equivalent Games in Extensive Form," *International Journal of Game Theory*, 26, 113–136.

Reutskaja, E., R. Nagel, C. F. Camerer, and A. Rangel (2011): "Search Dynamics in Consumer Choice under Time Pressure: An Eye-Tracking Study," *American Economic Review*, 101, 900–926.

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.

Salmon, T. C. (2004): "Evidence for Learning to Learn Behavior in Normal Form Games," *Theory and Decision*, 56, 367–404.

Schotter, A., K. Weigelt, and C. Wilson (1994): "A Laboratory Investigation of Multiperson Rationality and Presentation Effects," *Games and Economic Behavior*, 6, 445–468.

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.

Zauner, K. G. (1999): “A Payoff Uncertainty Explanation of Results in Experimental Centipede Games,” *Games and Economic Behavior*, 26, 157–185.