Tree construction and backward induction: a mobile experiment[*]

Konrad Grabiszewski[†] Alex Horenstein [‡]

Preliminary Draft
May 2018

Abstract

Game theory is a collection of tools to analyze interactive situations. We focus on two fundamental tools employed in the context of dynamic games: tree construction (modelling reality in its simplified form) and backward induction (solving the tree in terms of strategy profiles). We ask whether these tools work; that is, whether people behave as if they were constructing trees and backward inducting. We find that sometimes these tools work and proceed to explain when and why they fail. We find to key forces: subject's skills and complexity of interaction. We show how to measure skills and complexity. We also analyze the relationship between the structure of interaction and its complexity. Finally, we compare the relative importance of skills and complexity in terms of increasing the likelihood of subjects constructing a tree and backward inducting. In order to collect the data, we conduct a mobile experiment: we developed the mobile game *Blues and Reds* that is available for free for both iOS and Android devices.

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

[*] We express our gratitude 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.

[†] Department of Economics, University of Miami, USA; konrad.grabiszewski@gmail.com

[‡] Department of Economics, University of Miami, USA; alexhorenstein@gmail.com

1 Introduction

Game theory as a toolbox. For social scientists, game theory is, in the words of Osborne and Rubinstein (1994), "a bag of analytical tools designed to help us understand the phenomena that we observe when decision-makers interact." These tools – models and solution concepts – serve two roles. A model depicts how interacting decision-makers perceive the reality. A solution concept identifies what strategies these decision-makers select. In this paper, we use an innovative experimental methodology to offer a novel study of dynamic game theory and its two fundamental tools, *tree construction* (model) and *backward induction* (solution concept).

Does game theory work? We start with testing whether game theory works; that is, we ask two questions: *Do people construct trees?* and *Do people backward induct?* To clarify, we do not interpret these questions literally; rather, what we test is whether people "behave as if" constructing trees and backward inducting.

To be more precise, to test tree construction means to verify whether a tree correctly captures subject's perception of reality; if it does, then we say that the subject constructs a tree. To test backward induction means to check if backward induction correctly identifies the strategies that subject chooses; if it does, then we say that the subject backward inducts.

We find that game theory works for some subjects in some interactions but sometimes game theory does not work. What causes that "sometimes" is what we focus on next.

Why does game theory not work? Our tests of tree construction and backward induction indicate not only that these tools do not work but also allow us to discover two forces which explain why game theory fails: subjects and interactions.

First, we observe that the relevant strategic skills are not homogenous across our sample. This explains why game theory sometimes fails: some subjects have skills too low to construct a tree or to backward induct. Since skills are not directly observable, our challenge is to identify how to measure skills.

Second, we find that interactions differ in their complexity which also explains why game theory sometimes fails: some interactions are too difficult for a subject to translate them into trees or to solve using backward induction. We identify how to empirically measure the complexity of interaction and also determine how the structure of interaction affects its complexity.

In our final empirical exercise, we address the question of what — subject's skills or complexity of interaction — matters relatively more in the context of increasing the probability of game theory working (i.e., of people constructing trees and backward inducting). It turns out that for both, tree construction and backward induction, it is the skills that matter more.[1]

Mobile experiment. In order to collect the data, we decided to take an unconventional but innovative route: 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. For the purposes of this paper, we developed *Blues and Reds*, a mobile game available globally for free for iOS and Android devices.[2] Our mobile game consists of 58 levels (interactive situations) played against Artificial Intelligence (AI). Each level is a two-person, turn-based, zero-sum finite game with perfect and complete information. In order to facilitate the discussion of our paper, we would like to invite the readers to play *Blues and Reds.*

Structure of the paper. Tree construction and backward induction are two separate game-theoretic tools serving two separate roles. Consequently, it would seem natural to study them in 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 analysis of dynamic game theory, we merge two studies in one paper.

In section 2, we place our paper in the literature. In section 3, we present *Blues and Reds* as an experiment. In Section 4, we find that game theory fails and, in section 5, we explain why this is the case. Section 6 concludes.

[1] Our skills-versus-complexity analysis is important from the perspective of designing dynamic games. For instance, consider a regulator who proposes a new law that aims at modifying the behavior of interacting agents. Using game theory, the regulator theoretically predicts what is going to happen. Given that prediction, the proposed regulation becomes the reality. However, the observed outcome differs. As our results indicate, this is because the agents' skills are too low or the regulation as an interactive problem is too complex. In order to fix the failed regulation, the regulator wants to spend resources but must choose between improving agents' skills (e.g., training) or making regulation less complex (e.g., re-design/simplification of law). The choice crucially depends on what (skills or complexity) matters relatively more; our study helps with this choice: education is more important.

[2] 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.

2 Our paper in the literature

Although perception (i.e., tree construction) takes place before the selection of strategy (i.e., backward induction), our empirical analysis in sections 4 and 5 goes in the reverse order: first, we look at backward induction, then we study tree construction. This is because the empirical strategy we deploy for the study of tree construction requires a prior analysis of backward induction.

In order to make the structure of this section consistent with the structure of sections 4 and 5, first, we review the literature on backward induction and what we add to this literature (section 2.1); second, we discuss the literature on tree construction and our contributions to this literature (section 2.2). Finally, we also address the methodological aspects of our study (section 2.3).

2.1 Backward induction

Although fundamental in theoretical and applied economics, backward induction does not survive the stress of empirical testing. Thousands of papers have confirmed that, in general, people do not backward induction. The relevant question is not whether people backward induct but why they do not backward induct. The literature that attempts to explain the discrepancy between the behavior that we observe and the behavior that game theory predicts can be divided into two themes: "blame the subject" and "blame the researcher."

We contribute by discovering two new forces which explain the violations of backward induction. We also analyze which force is relatively more important; as far as we known, there is no similar comparative study in the literature. Our first explanatory force expands the list of ways to "blame the subject." The second force opens a novel theme, "blame the interaction."

Blame the subject. According to this stream of literature, the observed discrepancy between theory and data is due to some imperfection on the part of a subject.

1. Imperfect strategic reasoning. The fundamental model of imperfect strategic reasoning is the level-k model.[3] This model has been extensively tested in the literature and the main message is

[3] 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).

that people, indeed, struggle with strategic reasoning.[4]

2. Imperfect cognitive skills. As of cognitive skills, it has been established that these skills are correlated with the behavior being more consistent with the theory.[5] In addition, we also know that subject's cognitive skills also influence the behavior of other subjects who take into account limited abilities of their opponents.[6]

3. Imperfect perception. The problem of perception is the problem of creating a model of interaction (i.e., tree construction) which we also analyze in this paper; we delegate the discussion of the relevant literature to section 2.2.

4. Imperfect strategic skills. To the list of imperfections, we add a new one: subject's strategic skills. While it is obvious that skills matter, what is less obvious is how to measure these skills. Our contribution is a novel, two-layer measure that allows to rank people in terms of how likely they are to backward induct. Our approach is based on the classical problem of economics: allocation of limited resource. In our experiment, this limited resource is time which subjects allocate across different rounds of a tree. By looking at the subject's time allocations, we are able to determine the level of subject's strategic skills and, consequently, the likelihood of a subject backward inducting.

Blame the researcher. According to this stream of literature, the discrepancy between theory and data is due to the researcher's assuming incorrect model of interaction. That is, the subjects only seemingly violate backward induction. Experiments following this approach do not test backward induction alone but rather a joint hypothesis: occurrence of some phenomenon *and* subjects backward inducting. Consequently, the behavior observed in these experiments should not be interpreted as a rejection of backward induction; rather, these experiments do not reject the joint hypothesis they test. This would be the case of experiments with the very popular ultimatum game and centipede game. Since these experiments test for more than just backward induction, they differ from the "blame the subject" literature and our paper.

[4] E.g., Ho et al. (1998), Costa-Gomes et al. (2001), Bosch-Domènech et al. (2002), Costa-Gomes and Crawford (2006), Costa-Gomes and Weizscker (2008), Wang et al. (2010), Agranov et al. (2012), Arad and Rubinstein (2012a), Burchardi and Penczynski (2014), Hargreaves Heap et al. (2014), Shapiro et al. (2014), Georganas et al. (2015), Fehr and Huck (2016), Penczynski (2016), Batzilis et al. (2017), and Ho and Su (2013). See also experimental studies related to level-k model in Kneeland (2015), Bayer and Renou (2016a), and Friedenberg et al. (2017).

[5] 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).

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

Blame the interaction. As far as we know, ours is the first paper to empirically find that the trees differ in their complexity. Complexity of interaction is yet another explanation of why backward induction is rejected: some trees are just too difficult for the subjects to solve. Since our experiment consists of a multiplicity of trees, we not only discover that complexity matters but also explain how a tree structure affects its complexity.[7]

2.2 Tree construction

The literature analyzing tree construction — that is, whether game theory correctly depicts how people perceive the reality — is very small, especially, in comparison to the vast literature on backward induction. And this is despite the fact that the problem of perception was already recognized in the literature 60 years ago when Simon (1957) pointed at the fact that cognitive effort is costly and might induce sub-optimality (in comparison to costless cognition).

From the small literature on perception, we know that people do not seem to be able to correctly determine the connection between what they choose and what they obtain.[8] When playing the same game but depicted in different, yet equivalent, forms (normal vs extensive) people change their behavior.[9] It is not very surprising that people do not correctly perceive the interactions because often they do not even search for the relevant information about the very interaction they participate in.[10]

As far as we know, ours is the first in-depth analysis of tree construction. We begin with showing that it is not always the case that people construct a tree. As with the study of backward induction, we find two explanations.

[7] McKelvey and Palfrey (1992) and Fey et al. (1996) experiment with the centipede game of different lengths; 4-move and 6-move in the former, and 6-move and 10-move in the latter. Not surprisingly, both find that backward induction is more likely to be violated in a longer centipede. Dufwenberg et al. (2010) and Gneezy et al. (2010) use two forms of the race game; one larger than the other. Each finds that the failure rate of backward induction is higher in a larger game. It also is important to note that Gill and Prowse (2017) also measure the complexity of interaction but they focus on static games (beauty contest) while our interest lies in dynamic games and their structure.

[8] E.g., Chou et al. (2009), Rydval et al. (2009), and Cason and Plott (2014).

[9] E.g., Schotter et al. (1994), Rapoport (1997), and McCabe et al. (2000). Cox and James (2012) show — in the context of the Dutch auction and the centipede game — that behavior depends on the specific format in which the game is presented.

[10] 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).

First, we “blame the subject”: some people do not posses high enough skills. We show how to measure these skills. Second, we “blame the interaction”: some interactions are too complex for people to translate them into trees. We also investigate what causes one interaction to be more complex than another interaction. Finally, we also analyze the relative importance of skills versus complexity.

Given how extensive the empirical literature on backward induction is, we do not need to convince anyone that studying backward induction is a valuable exercise. However, very little has been said about tree construction; hence, we feel the need to emphasise the importance of testing whether people analyze interactions 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 theoretical decision-makers who have the privilege of dealing with trees as it is only in academic papers textbooks that the interactive situations are depicted as trees. In reality, people participate in interactions depicted as anything but trees (e.g., verbal representation, alternative graphical representation).

Creating models of reality, not to mention constructing trees, is neither trivial nor natural task. Game theory offers no help in tree construction. This is not surprising as a tree captures its creator’s perception and, consequently, tree construction is more of an art than science. A subject’s model of reality might have nothing to do with how an experimenter represents the very same reality. On the other hand, game theory has a lot to say when it comes to the choice of strategies in a given model. Solution concepts are more of science than art.

Since we should not assume that people analyze interactive situations as game theory does, neglecting the question of whether people perceive interactive situations in accordance with theory has serious consequences. If people do not construct trees, then testing backward induction is a meaningless exercise that might lead to erroneous conclusions. First, suppose that a subject impeccably backward inducts when solving a tree. Is it enough to confidently forecast that the subject’s choices would be consistent with backward induction when she deals with non-tree interactions (external validity)? The answer is negative since it is likely that the subject lacks the relevant skills to analyze those non-tree interactions. Without such skills, it impossible for her to make the right choice.

Alternatively, suppose that when it comes to the situations in non-tree forms, our subject does not choose in line with backward induction. Can we conclude that the subject does not backward induct (internal validity)? We cannot because it is possible that that subject applied backward induction but on the incorrect tree (or she maybe even did not construct a tree at all).

To sum, while neglected by the literature, testing tree construction is important because real-life interactions are not depicted as trees, we have no reason to assume that people construct trees when analyzing the interactions, and without knowing that people construct trees, there is little value in testing backward induction.

2.3 Methodology

With its non-standard way of collecting data, our paper belongs to the line of research that relies on innovative methods to gather the data.[11] 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. Obviously, this openness is not possible in the more standard setup (lab). Other innovative experimental methodologies include the use of newspapers (from the pre-Internet era; for examples, see Bosch-Domènech et al. (2002)) and web-based experiments (e.g., Ariel Rubinstein's `gametheory.tau.ac.il`).

In our design of measuring skills and complexity, we rely on subjects' response times (RT), a tool that just 15 years ago was rather unlikely to be seen in an economics paper. We believe that the change in the economists' attitude towards RT is due to the 2004 Presidential Address to the Econometric Society by Ariel Rubinstein (Rubinstein (2006)). Since then, the empirical literature using RTs has been growing very fast.[12]

[11] As far as we know, we are the first economists who created a mobile game in order to collect the experimental data.

[12] For the reviews of the relevant literature, see Clithero (2016) and Spiliopoulos and Ortmann (2017).

3 Experimental Design

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

Levels in *Blues and Reds* are divided into two types: tree and non-tree (as in Cox and James (2012)). Tree levels, 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. Figure 1, a screenshot from *Blues and Reds*, is an example of a tree level.

[Figure 1 about here.]

Non-tree levels, 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 from subjects to spend effort. Figure 2, another screenshot from *Blues and Reds*, is an example of a non-tree level.

[Figure 2 about here.]

An important element of our experimental design is that for each tree level there is an equivalent non-tree level. Equivalency means that a tree level and the tree that is an extensive form of the equivalent non-tree level are identical. For example, Figures 1 and 2 are equivalent levels as the former is a tree representation of the latter.

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

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 level, 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 the issues like payoff uncertainty Zauner (1999)) or experimenter's misunderstanding of subject's utility function.

Every level is winnable; that is, a subject can win. However, to win the subject must not make a mistake. Starting at the initial node, there is only one action that would lead to the subject's win. This no-mistake property holds at the subsequent nodes 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 minimize 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 mental model or backward inducting.

Second, no mistake-property removes the (ir)rationality of opponent as a factor affecting behavior of our subjects. As expected theoretically and confirmed empirically (e.g., Palacios-Huerta and Volij (2009), Agranov et al. (2012), Alaoui and Penta (2016), Fehr and Huck (2016), and Gill and Prowse (2016)), what a subject thinks about rationality of her opponents affects the subject's behavior. However, in our experiment, it does not matter what a 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 would be zero.

Every tree and non-tree level has a 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. For example, Figures 1 and 2 are 3.2.3 levels. 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 list all 27 levels, sorted by the number of rounds, that we use to build our data.

[Table 1 about here.]

The decision to consider only interactions with the structure $N_1.N_2.N_3.N_4.N_5.N_6$ is motivated by one of our research questions: we want to understand how the structure of interaction (of a tree in the case of tree level, or of the equivalent tree in the case of non-tree level) drives the observed behavior. This, in particular, requires that every path (from the initial node to a final node) has the same length.

To elaborate, consider centipedes in Figures 3 and 4, respectively. In each tree, a quick glance at the payoffs indicates that the subject should play β_1 at her first node. There is no need for the subject to exert effort on further analysis. It does not matter that Tree 2 is bigger than Tree 1 — we expect the same percentage of wins and loses in each tree. Consequently, we learn nothing about the impact of tree structure on behavior.

[Figure 3 about here.]

[Figure 4 about here.]

Another important aspect of our mobile game is the “one life per level” feature: except for the tutorial levels, a subject can play each level only once. If she wins a level, then she collects a star (an in-game reward) but can not play that level again. If she loses, then she does not gain a star and is unable to try again. While this feature is very uncommon in mobile game, we introduced it in order to motivate subjects to think rather than mindlessly select their choices. Subjects are informed about the “one life per level” feature in the tutorial as well as reminded about it right before they start a new level (“Remember, in each level you have only one life!”)

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 over 6,000 subjects who played at least one level of *Blues and Reds* beyond the mandatory tutorial. For each level and each subject, we collected the following data.

1. Whether subject wins or loses.

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

Since RT is a fundamental metric in our study, we removed observations above the 95th percentile of RT within each level to avoid outliers generating an upward bias. This is so because some players might have stop playing a certain level for multiple reasons and come back to play at a much later time[Footnote: results remain qualitatively the same if instead of trimming the sample at the 95th percentile of RT for every level we do it at the 90th percentile]. For example, one of the outliers removed is a player that took more than 10 days to make a move.

Our final sample consist of 6,677 players who have played 44,113 trees and spent less than 533 seconds solving a tree.

We use tree levels to study backward induction. Given the properties of tree levels, we say that a subject who won in a tree was backward inducting in that tree.

We use pairs of equivalent tree and non-tree levels to study tree construction. Given the properties of non-tree levels, 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, we are unable 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 levels which serve as our screening tests. For each tree level, we select subjects who win; these are our BI subjects. In the second step, we turn to non-tree levels. In a given non-tree, we restrict our attention to BI subjects from the equivalent tree level. If a BI subject wins in a non-tree, then we say that she constructed a tree.

We interpret response times as outcomes of the traditional economic choice problem: selecting how to use a scarce and costly resource (time). We analyze two aspects of time, how much time subjects spend and how they spend that time across rounds.

4 Does game theory work?

Our empirical analysis begins with the obligatory analysis of whether game theory works; that is, whether people backward induct and construct trees.

4.1 Backward induction

As explained in section 3, a subject who loses in a tree level did not backward induct in that tree. For each of 27 trees, we compute the failure rate of backward induction; i.e., the percentage of subjects who did not backward induct. That failure rate ranges from 2.5% to 53.1% (Table 2). Theory predicts that we should observe 0% in each tree.

[Table 2 about here.]

Table 2 not only confirms what is already known in the literature (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 the relevant strategic skills. If subjects were identical or their skills were not relevant, then all subjects would perform equally well or equally poorly and 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 induct would be the same in each tree.

4.2 Tree construction

As a explained in section 3, a subject who wins in a tree level but loses in the equivalent non-tree level did not construct a tree in that non-tree. For each of 27 non-trees, we compute the failure rate of tree construction; i.e., the percentage of backward inducting (BI) subjects who did not construct a tree. That failure rate ranges from 12.98% to 63.98% (Table 3) while theory assumes that the failure rates should be zero.

[Table 3 about here.]

As in the case of backward induction, we observe that, in general, subjects do not construct trees. We also note that non-trees differ in complexity (because the failure rates differ) as well as subjects' differ in the relevant tree-construction 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.

5 Why doesn't game theory work?

In section 4, we learn that game theory does not work: our subjects neither backward induct nor construct trees. We argue that these violations of game theory are driven by skills of our subjects (being too low) and complexity of the games they played (being too high). In this section, we study why backward induction and tree construction fail.

5.1 Backward induction

5.1.1 Subject's skills

People differ in their abilities to backward induct. Higher skills imply higher probability of a subject backward inducting. However, skills are not directly observable. What is observable is the choices which reflect those skills. Our objective is to figure out how to measure the strategic skills.

What we observe in our data is response time (RT, measured in seconds); that is, the time that a subject spends at a given round of a multi-round tree. We think of time as a scarce resource and subjects allocating that resource across various rounds of our dynamic games.

In the context of static games, there is only one response time, the total time that a subject spends choosing a strategy, as there is only one instance when a subject makes a choice. Hence, to measure skills or type, researchers look at the total time subjects spent on solving the game as this is the only RT that we measure.[13]

This is not the case of dynamic games which makes our analysis of response times more complicated. To elaborate on the problem of designing RT-based measure of skills in dynamic games, consider

[13] E.g., Rubinstein (2006), Rubinstein (2007), Rubinstein (2013), Rubinstein (2016), Arad and Rubinstein (2012b), and Gill and Prowse (2017).

a 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.

[Table 4 about here.]

According to the total time spent solving the tree, 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 time spent on the first node, Chris is the fastest thinker followed by Ann, David, and Bob. Finally, Ann and Bob allocate 75% of the total time at the first node while Chris and David allocate only 40%. How will the subjects in Table 4 rank in terms of their likelihood to backward induct in 2.2.4.2?

Intuition might indicate that it is the total time spent on solving a tree or just the time spent at the initial node (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 or time spent at the first node 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 (total and at the first node) and is more likely to backward induct compared to someone who has never heard of game theory.

In order to measure skills relevant for backward induction, we develop a two-layer measure. We find that what matters most is not necessarily how much time a subject spends but how she spends that time; the total time spent on shoving a tree is of secondary importance. That is, the allocation of limited resource is more important than the amount of that resource.

The key measure is the relative time spent at the beginning of interaction; that is, the time spent at the first node as the percentage of total time spent. The higher that measure the more likely a subject is to backward induct. We explain our result in the following way. When it comes to backward induction, what matters is the time spent at the first node. 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 had already designed. This is, especially, true in our trees where making a mistake at any round is equivalent with losing a level.

In fact, if a subject correctly had analyzed the tree at the very beginning, then she does analyze what to choose in the subsequent rounds and what is needs the time for (in our experiment) is only to physically select an action. However, if a subject also spends significant amount of time at the following rounds, then it means she doubts herself or made a mistake in a previous round. This tells us that the higher the relative time at the first node the more likely a subject is to backward induct.

In order to prove that the relative time spent at the first node is the key measure, we conduct the following empirical exercise. We consider three candidates for the measure of skills.

- TT. Total time that a subject spent on solving a tree. It is the sum of RTs from all rounds.
- T1N. Time that a subject spent on the first node. It is RT from the first round.
- RT1N. Relative time that a subject spent on the first node as a percentage of total time. This is T1N as percentage of TT.

For each tree, we divided the subjects in terciles[14] by the corresponding measure. For example, take TT. The Low tercile corresponds to all subjects having TT less or equal to the percentile 33.3%. The Medium tercile corresponds to the subjects having TT higher than the 33.3% percentile and less or equal than the 66.6% percentile. Finally, the High TT tercile corresponds to players with TT higher than the 66.6% percentile. Since some players have equal values of TT, T1N, and RT1N, each tercile does not necessary contains the exact same number of players.

For each tercile, we compute the percentage of subjects in that tercile that backward inducted. For instance, for the tree 2.2.2 and measure R1TN, 84%, 98%, and 99% of subjects in the Low, Medium, and High tercile, respectively, backward inducted. The results are presented in Table 5. Since we are interested in relative time, our analysis excludes 2-round trees where RT1N = 100% by default.

[Table 5 about here.]

In Table 5, we observe that RT1N is the best predictor of subjects backward inducting. This measure never fails in a sense that for each tree, higher RT1N is associated with higher probability of backward inducting.

[14] Our conclusions remain unchanged if instead of terciles we use deciles.

As of TT, we have trees like, for instance, 2.2.2 or 2.2.2.2 where higher TT implies a subject being less likely to backward induct. However, for trees 3.2.2.2.2, 4.2.2.2.2, and 2.2.2.2.2.2, we have the opposite relationship: higher TT makes backward induction more likely. In addition, for trees 2.2.2.4, 4.2.2.2, 3.2.3, 3.2.2.2, 2.2.2.3, and 2.4.2.2 the relationship between TT and 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 T1N. In 2.2.4.2 or 2.2.2.2, higher T1N makes backward induction more likely. However, in 3.3.2, 3.3.3, 2.3.2, and 2.3.2, we again have a non-monotonic relationship.

According to RT1N, 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 RT1N. Is there any way to refine our ranking and compare Ann vs Bob as well as Chris vs David? It turns out there is.

We find that while TT does not allow us to rank the subjects by itself, it does help once we condition on RT1N: for a given RT1N, 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 of RT1N, first, we divide the observations of tree separately into terciles by RT1N. Then, each RT1N tercile of every tree is divided into another three terciles, this time, with respect to TT. Therefore, for each tree we have 9 pairs $(TT_i|RT1N_j)$ 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 player moves at least twice.

[Table 6 about here.]

We observe that for each RT1N tercile, increasing TT decreases the proportion of subjects who backward induct. Importantly, Table 7 shows that this result also holds for individual trees.

[Table 7 about here.]

We believe that our result about TT — given RT1N, TT ranks subjects — can be explained in a simple way. Consider Ann and Bob from Table 4. Both spend 75% of total time at the first node 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. This fast-slow difference is captured by TT.[15]

To summarize, when it comes to measuring skills relevant for backward induction, we find the following. The higher the relative time spent at the 1st node (RT1N), the more likely a subject is to backward induct. In addition, for a given value of RT1N, the higher the total time spent on solving the tree (TT), the less likely a subject is to backward induct.

5.1.2 Complexity of interaction

As our results in section 4, trees vary in their complexity. Our first objective is to design the empirical measure of complexity. Next, we want understand how the tree structure affects its empirical complexity.

Our design of the empirical measure of complexity begins with the observation that, in some cases, comparing trees requires no empirical work. One tree can be objectively less complex than another tree. To elaborate, consider the three trees depicted below: 2.2 (Figure 5), 3.3 (Figure 6), and 2.2.2 (Figure 7) with payoffs in order $(subject, AI)$. In each tree, subject moves at odd rounds and AI chooses at even rounds.

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

Tree 2.2 is objectively less complex than both trees 3.3 and 2.2.2 as it can be embedded in each of them. Formally, we define "objective complexity" in the following way. 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

[15] See Kahneman (2013), Rubinstein (2007), and Rubinstein (2016) for the analysis of slow/fast and contemplative/instinctive thinking. It is also possible that Bob is over-thinking (see Gill and Prowse (2017)) but this is not a hypothesis we can test with our data.

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.

When tree X is objectively more complex than tree Y, then we denote it by $X \trianglerighteq Y$. According to our definition of "objectively more complex," 2.2.2 is objectively more complex than 2.2, and 2.4.2 is objectively more complex than 3.2. However, 2.3 is not objectively less/more complex than 3.2.2 nor is 3.3 objectively more/less complex than 3.2.3. Our main focus is to identify and explain the non-objective comparisons.

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 Figure 8.

[Figure 8 about here.]

The main problem with objective complexity is that it does not generate complete ranking of trees. For example, it is not clear which tree — 3.3 (Figure 6) or 2.2.2 (Figure 7) — is more complex. This is the reason we need an empirical measure. 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).

It might seem that the percentage of subjects who fail to backward induct (i.e., failure rate of backward induction) would be the appropriate empirical measure of complexity. However, as we show, this is not the case. This is because winning/losing is a binary variable. As such, losing is not very informative. For example, a subject might lose in both 3.3 and 2.2.2 but it does not mean that these two trees are equally difficult.

For that reason we need a more refined variable; namely, the response time (RT). We already explained the interpretation of RTs in our experiment. In Table 8, we rank the trees in accordance to the three potential empirical measures of complexity.

- % NOT BI. Percentage of subjects who did not backward induct (it is the repetition of the third column in Table 2)

- ATT. Average total time that the subjects spent solving a tree.

- AT1N. Average time that the subjects spent at the first node.

[Table 8 about here.]

In order to analyze which empirical measure is the best extension of the objective measure, we conduct the following exercise. If a tree A is objectively more complex than tree B, then we check whether an empirical measure of complexity of tree A is higher than that of tree B at the 1% level of significance using a one-side t-statistic. If this is the case, then we say that the empirical measure agrees with the objective measure. If a tree A is objectively more complex than tree B but the empirical measure tells than it is B that is more complex at the 1% level of significance, then we say that the empirical measure and objective measure disagree. Finally, if A is objectively more complex than B but their empirical complexities do not differ at 1%, then we say that the result of comparison is undefined. The results are depicted in Table 9.

[Table 9 about here.]

Looking at the Table 9, we conclude that ATT and AT1N are very similar and far superior to % NOT BI. In fact, ATT and AT1N almost perfectly replicate the objective ranking of complexity. Consequently, we consider both ATT and AT1N as the appropriate empirical measures of complexity of backward induction. In the rest of this section, we present the results only for AT1N as our measure of complexity; qualitative results with ATT as a measure of complexity are the same.

To complement Table 9, we present the heat map in Figure 9 which we read in the following way. Take a tree X from the horizontal axis and a tree Y from the vertical axis. Tree Y is empirically more complex than tree X, and the color indicates how significant is the difference in the empirical complexities between Y and X. Gray means that the difference is not statistically significant, yellow means significant at the 10% level, orange means significant at the 5% level, and red means significant at the 1% level or less.

[Figure 9 about here.]

Our next objective is to analyze the main drivers of empirical complexity. In particular, we find connections between the tree structure and empirical complexity. When we start with a given tree, then we can add more rounds (make the tree longer) or more actions at the already existing non-final nodes (make the tree wider). The three questions we ask are the following:

1. Is longer tree more complex?
2. Is wider tree more complex?
3. What is increasing complexity more: making a tree longer or wider?

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 10 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 10 clearly shows that increasing the number of rounds increases the empirical complexity of a tree.

[Figure 10 about here.]

We now study the impact of increasing the number of actions per round. The analysis is presented in Table 10. First, we partition trees according to the number of rounds. Second, each subset of trees with the same number of rounds, we divide into three group — 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.

[Table 10 about here.]

To understand the impact of the width on tree complexity, we look at Panel (b) in Table 10 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 tree wider makes it more complex.

Finally, we analyze what makes a tree more complex: making it longer or making it wider. To that end, we propose the following exercise. Whenever we elongate or widen a give tree, the number of final nodes increases. That number is important as it represents the number of paths that a decision-maker must analyze. In our exercise, we take a tree and expand it in two directions — length and width — but in such a way that the increase of final nodes is the same for both elongating and widening. In our data, there are three cases of such an analysis.

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, empirical complexity increases to 13.53. Alternatively, we can make 2.2 wider by expanding it to 2.4 (8 final nodes); 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 empirical complexity 21.61. Alternatively, we can make 2.2.2 wider by expanding it to 4.2.2 (16 final nodes) with empirical complexity 17.31. Here, 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 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 29.95. Here, we observe that elongation is more important.

To summarize, we observe that in each case it is the length that has a bigger impact on complexity.

5.1.3 Skills vs complexity

So far, we discovered how to measure subject's skills and complexity of tree. We also learned that both skills and complexity have an impact on whether a subject backward inducts. Here, we analyze which factor, skills or complexity, is relatively more important. To that end, we run a logit regression to assess the impact of strategic skills (Relative Time) and tree complexity (Average Total Time) on the probability of subject backward inducting. More precisely, we use the Maximum Likelihood Estimation (MLE) to estimate the parameters of the following standard logit model.

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

Y is the vector of dependent variables in the regression capturing whether the player backward inducted ($Y = 1$) or did not backward induct ($Y = 0$), α is the intercept, and X is the $N \times 3$ matrix of independent variables: Skills (relative time at the first node; RT1N), Complexity (Average Total Time; ATT), and Sequence. The independent variable Sequence is the order in which a tree appeared in a subject's sequence of trees (from 1 to 27) and N is the number of observations in the sample. Table 11 depicts summary statistics of the independent variables.

[Table 11 about here.]

Table 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. Third, using the cutoff value of 50%, the model shows good predictive power: it correctly predicts whether a subject is going to backward induct or not in 86% of the cases.

[Table 12 about here.]

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 13.

[Table 13 about here.]

We interpret the results in Table Table 13 in the following way. Increasing subject's kills by one standard deviation increases probability of backward induction by 19.97%. Increasing tree complexity by one standard deviation decreases probability of backward induction by 6.21%. Finally, if the tree appears in sequence T instead of $T - 1$, 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 bigger impact on the probability of a subject backward inducting.

5.2 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 in section 5.1. The only aspect that changes is the data. Recall that in order to test for tree construction, we look at the behavior in non-tree levels of subjects who backward inducted in the equivalent tree levels. We believe that this is a very attractive feature of our study.

Given that we only keep data from subjects who solved the equivalent tree level of the non-tree problem, our data in this section is smaller than that used in the previous one. At the same time, after removing the data in which a player did not solved the tree correctly, in this section we also removed observations above the 95th percentile of TT within each level to avoid outliers generating an upward bias. The final dataset used in our analysis consists of 4,646 players who played 28,587 non-tree levels after winning the corresponding tree level.

5.2.1 Skills of subjects

As it was the case of backward induction, people also differ in their abilities to construct trees. Again, our challenge is to design the measure of the relevant skills. We proceed as we did in section 5.1.1. In Table 14, we replicate the analysis we conducted for backward induction in Table 5. We observe that, as it was the case in our analysis of backward induction, the best measure to capture tree-construction skills is the relative time that a subject spent on the first node (RT1N).

[Table 14 about here.]

Next, we turn to analyze the conditional behavior of total time (TT). In Table 15, we first divide all the data into terciles by RT1N. Then, we divide each tercile into additional terciles with respect to TT. (This is a replication of Table 6.)

[Table 15 about here.]

We observe that for each RT1N tercile, increasing TT decreases the proportion of subjects who construct a tree. As Table 16 shows, this is true in each individual non-tree.

[Table 16 about here.]

To summarize, from the qualitative perspective, we measure backward-inducting skills and tree-constructing skills in the same way. The higher the relative time spent at the 1st node (RT1N), the more likely a subject is to construct a tree. In addition, for a given value of RT1N, the higher the total time spent on solving the non-tree (TT), the less likely a subject is to construct a tree.

5.2.2 Complexity of interaction

In order to analyze the complexity of non-trees, we follow the same strategy as we did in section 5.1.2. First, we design the empirical measure of complexity of constructing a tree. Next, we want understand how the structure of interaction (non-tree) affects its empirical complexity.

As with trees, we also want our measure of complexity to be the best extension of the objective measure. Following the same approach as in the study of backward induction, we start with Table 17 in which we rank the non-trees in accordance to the three potential empirical measures of complexity.

- % NOT TC. Percentage of subjects who did not construct a non-tree (it is the repetition of the third column in Table 3)
- ATT. Average total time that the subjects spent solving a non-tree.
- AT1N. Average time that the subjects spent at the first node.

[Table 17 about here.]

Next, in order to pick the best measure of complexity, in Table 18, we replicate the analysis from Table 9.

[Table 18 about here.]

Table 18 tells us we conclude that ATT and AT1N are the best measures of complexity. Hereafter, all the results are computed for AT1N but qualitative results remain the same if we use ATT as our measure of complexity. To complement Table 18, we present the heat map in Figure 11 which we read in the same way we read the heat map in Figure 9.

[Figure 11 about here.]

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 longer non-tree more complex?
2. Is wider non-tree more complex?
3. What is increasing complexity more: making a non-tree longer or wider?

First, we analyze the impact of length. Figure 12, which replicates Figure 10, 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. While computing average complexities for Figure 12, we removed the most complex non-tree; viz. 4.2.2.2.2. This is because its individual complexity is 983 (i.e., on average, subjects spend 983 second on the first node of this non-tree) while the complexity of second most complex non-tree (2.2.2.2.2) is 184.

[Figure 12 about here.]

Second, we analyze the impact of width. Table 19 replicates the analysis we conducted in Table 10. We observe that the results are not as clear as in the case of backward induction. For the interactions with two rounds, we notice that complexity initially increases to decrease for the widest interactions. In the case of 5-round interactions, we seem to have a U-shaped curve. Hence, we very cautiously conclude that wider interactions are more difficult for the subject to depict them as trees as the evidence is not as strong as we could have hoped for.

[Table 19 about here.]

Finally, we analyze what makes 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 trees. We look at three cases.

Case 1. We start with the non-tree 2.2 (4 final nodes) whose empirical complexity is 12.86. We can make 2.2 longer by expanding it to 2.2.2 (8 final nodes); with this, empirical complexity increases to 23.99. Alternatively, we can make 2.2 wider by expanding it to 2.4 (8 final nodes); empirical complexity increases to 13.89. 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 23.89. We can make 2.2.2 longer by expanding it to 2.2.2.2 (16 final nodes) with empirical complexity 36.64. Alternatively, we can make 2.2.2 wider by expanding it to 4.2.2 (16 final nodes) with empirical complexity 38.19. Here, we observe that making an interaction wider is more important.

Case 3. We start with the non-tree 2.2.2.2 (16 final nodes) whose empirical complexity is 36.64. We can make 2.2.2.2 longer by expanding it to 2.2.2.2.2 (32 final nodes) with empirical complexity 184.32. 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 47.22. Here, we observe that elongation is more important.

In two cases, we observe that it is the length that has a bigger impact on complexity. In one case, it is the width that is the dominant force. The final conclusion is not as clear as it was the case of backward induction.

5.2.3 Complexity vs skills

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

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

Y is the vector of dependent variables in the regression capturing whether the player created a tree ($Y = 1$) or did not create a tree ($Y = 0$), α is the intercept, and X is the $N \times 3$ matrix of independent variables: Skills (relative time at the first node; RT1N), Complexity (Average Total Time; ATT), and Sequence. The independent variable Sequence is the order in which a non-tree appeared in a subject's sequence of non-trees (from 1 to 27) and N is the number of observations in the sample. Table 20 depicts summary statistics of the independent variables.

[Table 20 about here.]

Table 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) cutoff value of 50%, the model shows good predictive power: it correctly predicts whether a subject is going to backward induct or not in 86% of the cases.

[Table 21 about here.]

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 22.

[Table 22 about here.]

Increasing subject's skills by one standard deviation increases the probability of tree construction by 31.69%. Increasing tree complexity by one standard deviation decreases that probability by 2.65%. Finally, if a non-tree appears in sequence T instead of $T-1$, the probability of tree construction increases by 085%. The comparison of relative importance between skills and complexity indicates that it is the skills that have bigger impact on a subject constructing a tree; in fact, a bigger impact that it is the case of backward induction.

6 Conclusions

Does game theory work? or rather *Why doesn't game theory work?* is the fundamental question of the game-theoretic empirical literature. In this paper, we contribute to this literature by studying tree construction and backward induction. We find that people only sometimes behave in accordance with the theory.

We identify that subject's skills and complexity of interaction are the key driving forces behind the failure of game theory. Using observable data (response times), we determine how to measure skills and complexity. When it comes to the skills, for both tree construction and backward induction, it is the relative time at the first node that is the key metric. The higher that relative time the more likely a subject is to construct a tree and backward induct. In addition, conditional on the relative time at the first node, the higher the total time the less likely a subject is to create a tree and backward induct.

As of complexity, we find that for both tree and non-trees, it is the average total time or average time at the first node that is the best measure. We also establish that making an interaction longer or wider makes it more complex; i.e., a subject is less likely to create a tree or backward induct. We also find that the length is a more important dimension; this result is strong in the case of backward induction but not so much for tree construction.

Finally, we compare the relative importance of skills and complexity on tree construction and backward induction. In each case, we find that the skills have a bigger impact. Consequently, if the goal is to make game theory work, then improving skills is of primary importance while simplification of complexity a secondary one.

Our data comes from a mobile experiment. We create a mobile game *Blues and Reds* that has been globally available since August 2017 on iOS and Android devices.

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 (2012a): "The 1120 Money Request Game: A Level-k Reasoning Study," *American Economic Review*, 102, 3561–3573.

——— (2012b): "Multi-dimensional iterative reasoning in action: The case of the Colonel Blotto game," *Journal of Economic Behavior & Organization*, 84, 571–585.

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 job attachment," *Proceedings of the National Academy of Sciences*, 106, 7745–7750.

Burnham, T. C., D. Cesarini, M. Johannesson, P. Lichtenstein, and B. Wallace (2009): "Higher cognitive ability is associated with lower entries in a p-beauty contest," *Journal of Economic Behavior & Organization*, 72, 171–175.

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

Carpenter, J., M. Graham, and J. Wolf (2013): "Cognitive ability and strategic sophistication," *Games and Economic Behavior*, 80, 115–130.

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.

Chou, E., M. McConnell, R. Nagel, and C. Plott (2009): "The Control of Game Form Recognition in Experiments: Understanding Dominant Strategy Failures in a Simple Two Person "Guessing" Game," *Experimental Economics*, 12, 159–79.

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. Weizscker (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.

Dufwenberg, M., R. Sundaram, and D. J. Butler (2010): "Epiphany in the Game of 21," *Journal of Economic Behavior & Organization*, 75, 132–143.

Fehr, D. and S. Huck (2016): "Who knows it is a game? On strategic awareness and cognitive ability," *Experimental Economics*, 19, 713–726.

Fey, M., R. D. McKelvey, and T. R. Palfrey (1996): "An Experimental Study of Constant-sum Centipede Games," *International Journal of Game Theory*, 25, 269–287.

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*.

Gneezy, U., A. Rustichini, and A. Vostroknutov (2010): "Experience and insight in the Race game," *Journal of Economic Behavior & Organization*, 75, 144–155.

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.

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

KISS, H., I. RODRIGUEZ-LARAC, AND A. ROSA-GARCÍA (2016): "Think twice before running! Bank runs and cognitive abilities," *Journal of Behavioral and Experimental Economics*, 64, 12–19.

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.

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.

MCKELVEY, R. D. AND T. R. PALFREY (1992): "An Experimental Study of the Centipede Game," *Econometrica*, 60, 803–836.

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

OSBORNE, M. J. AND A. RUBINSTEIN (1994): *A Course in Game Theory*, MIT Press.

PALACIOS-HUERTA, I. AND O. VOLIJ (2009): "Field Centipedes," *American Economic Review*, 9, 1619–1635.

PENCZYNSKI, S. P. (2016): "Strategic thinking: The influence of the game," *Journal of Economic Behavior & Organization*, 128, 72–84.

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.

Simon, H. A. (1957): *Models of Man*, New York, NY: Wiley.

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.

List of Figures

Figure 1: Example of a tree level in *Blues and Reds*.

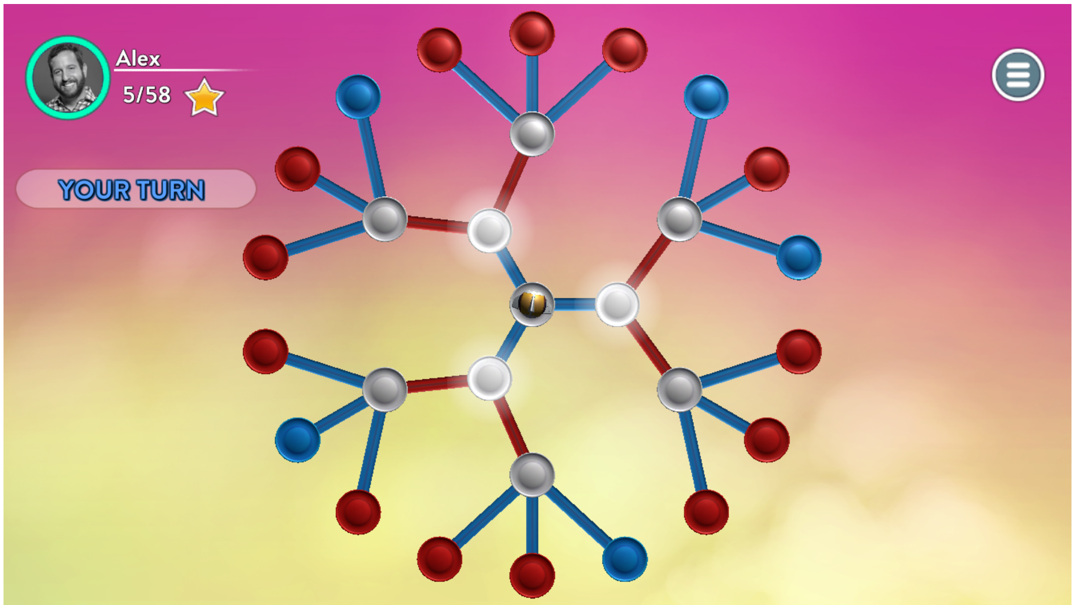


Figure 2: Example of a non-tree level in *Blues and Reds*.

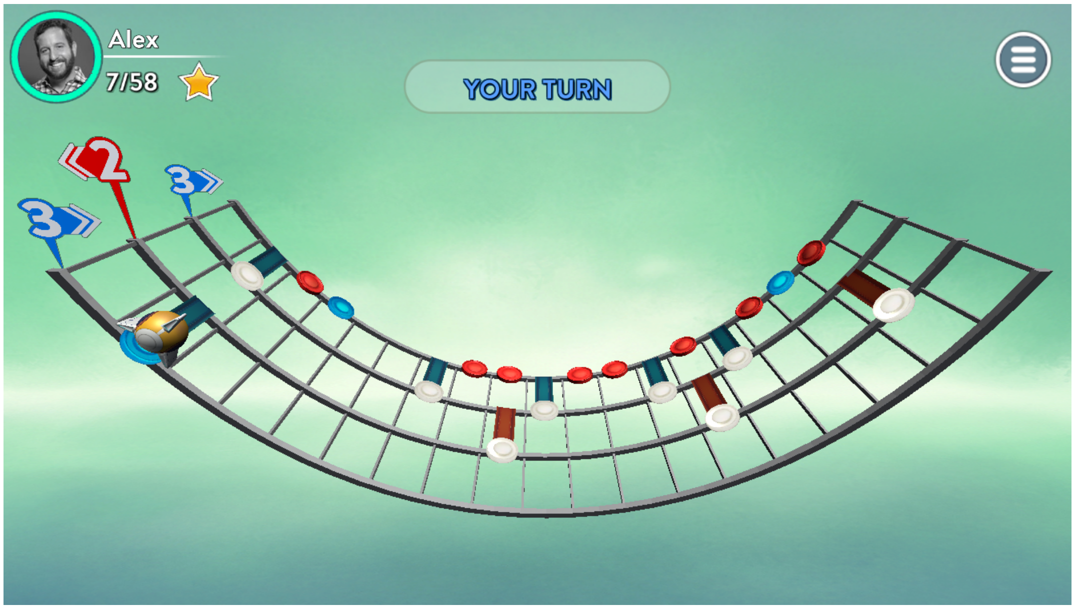


Figure 3: Tree 1

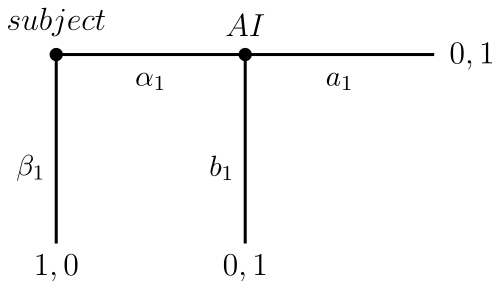


Figure 4: Tree 2

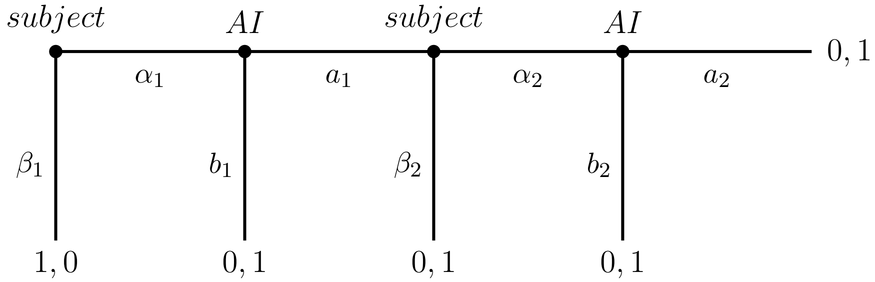


Figure 5: Tree 2.2

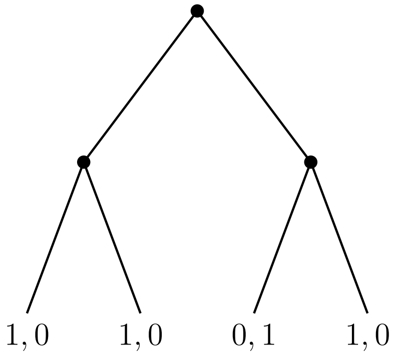


Figure 6: Tree 3.3

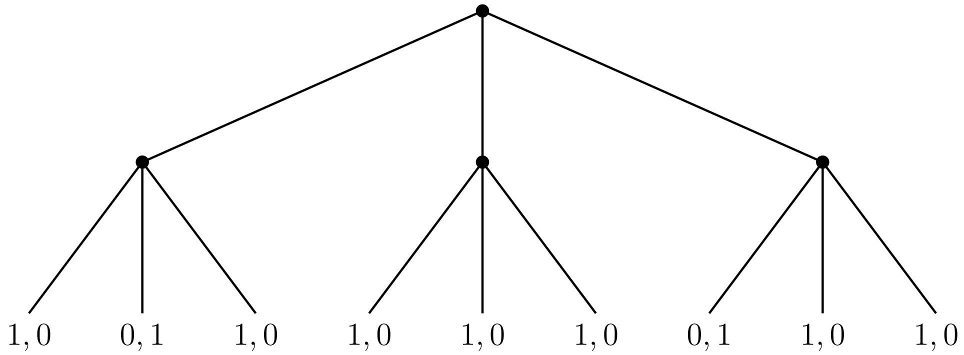


Figure 7: Tree 2.2.2

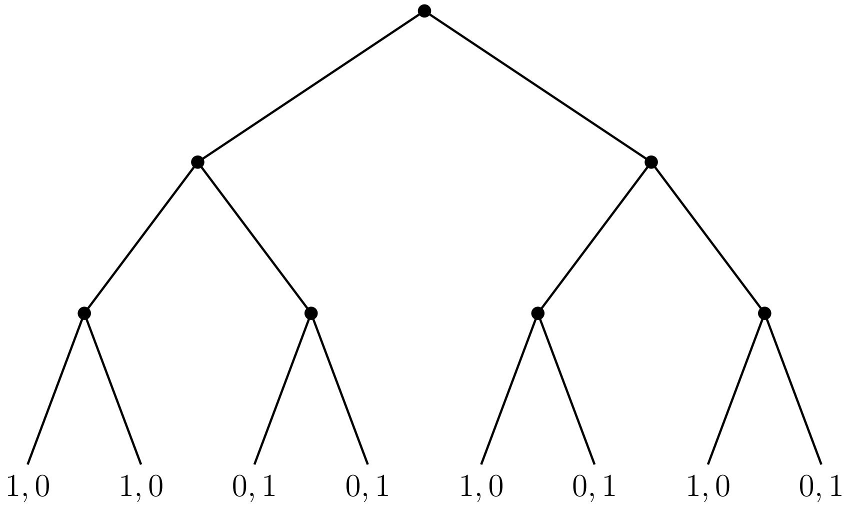


Figure 8: Pairwise comparison of trees in terms of objective complexity

This table presents the pairwise comparison of trees in accordance with objective complexity. Take a tree X from the horizontal axis and a tree Y from the vertical axis. If the symbol at the intersection of the Xth column and Yth row is $\trianglerighteq$, then $Y \trianglerighteq X$ (Y is objectively more complex than X), and if the symbol is a dot, then it is not possible to objectively compare X and Y.

	2.2	2.3	2.4	3.2	3.3	2.2.2	2.2.3	2.3.2	2.3.3	3.2.2	3.2.3	3.3.2	3.3.3	4.2.2	2.2.2.2	3.2.2.2	4.2.2.2	2.3.2.2	2.4.2.2	2.2.3.2	2.2.4.2	2.2.2.3	2.2.2.4	2.2.2.2.2	3.2.2.2.2	4.2.2.2.2	2.2.2.2.2.2
2.2.2.2.2.2	⊵	·	·	·	·	⊵	·	·	·	·	·	·	·	·	⊵	·	·	·	·	·	·	·	·	⊵	·	·	
4.2.2.2.2	⊵	·	·	⊵	·	⊵	·	·	·	⊵	·	·	·	⊵	⊵	⊵	⊵	·	·	·	·	·	·	⊵	⊵		
3.2.2.2.2	⊵	·	·	⊵	·	⊵	·	·	·	⊵	·	·	·	·	⊵	⊵	·	·	·	·	·	·	·	⊵			
2.2.2.2.2	⊵	·	·	·	·	⊵	·	·	·	·	·	·	·	·	⊵	·	·	·	·	·	·	·	·				
2.2.2.4	⊵	⊵	⊵	·	·	⊵	⊵	·	·	·	·	·	·	·	⊵	·	·	·	·	·	·	⊵					
2.2.2.3	⊵	⊵	·	·	·	⊵	⊵	·	·	·	·	·	·	·	⊵	·	·	·	·	·	·						
2.2.4.2	⊵	⊵	⊵	⊵	·	⊵	⊵	⊵	·	·	·	·	·	·	⊵	·	·	·	·	⊵							
2.2.3.2	⊵	⊵	·	⊵	·	⊵	⊵	⊵	·	·	·	·	·	·	⊵	·	·	·	·								
2.4.2.2	⊵	⊵	⊵	⊵	·	⊵	·	⊵	·	⊵	·	·	·	⊵	⊵	·	·	⊵									
2.3.2.2	⊵	⊵	·	⊵	·	⊵	·	⊵	·	⊵	·	·	·	·	⊵	·	·										
4.2.2.2	⊵	·	·	⊵	·	⊵	·	·	·	⊵	·	·	·	⊵	⊵	⊵											
3.2.2.2	⊵	·	·	⊵	·	⊵	·	·	·	⊵	·	·	·	·	⊵												
2.2.2.2	⊵	·	·	·	·	⊵	·	·	·	·	·	·	·	·													
4.2.2	⊵	·	·	⊵	·	⊵	·	·	·	⊵	·	·	·														
3.3.3	⊵	⊵	·	⊵	⊵	⊵	⊵	⊵	⊵	⊵	⊵	⊵															
3.3.2	⊵	⊵	·	⊵	⊵	⊵	·	⊵	·	⊵	·																
3.2.3	⊵	⊵	·	⊵	·	⊵	⊵	·	·	⊵																	
3.2.2	⊵	·	·	⊵	·	⊵	·	·	·																		
2.3.3	⊵	⊵	·	⊵	⊵	⊵	⊵	⊵																			
2.3.2	⊵	⊵	·	⊵	·	⊵	·																				
2.2.3	⊵	⊵	·	·	·	⊵																					
2.2.2	⊵	·	·	·	·																						
3.3	⊵	⊵	·	⊵																							
3.2	⊵	·	·																								
2.4	⊵	⊵																									
2.3	⊵																										
2.2																											

Figure 9: Heat map of empirical complexity (backward induction).

This figure, in a form of a heat map, presents the pairwise comparison of trees in accordance with their empirical complexity (average time that the subjects spent at the first node). Trees are ranked according to their empirical complexity from the easiest (2.2) to the most difficult (4.2.2.2.2). Colors indicate statistical significance of difference in empirical complexities. Take a tree X from the horizontal axis and a tree Y from the vertical axis. Red – Y is more complex than X at the 1% significance level, orange – Y is more complex than X at the 5% significance level, yellow – Y is more complex than X at the 10% significance level, and grey – Y is not statistically more complex than X.

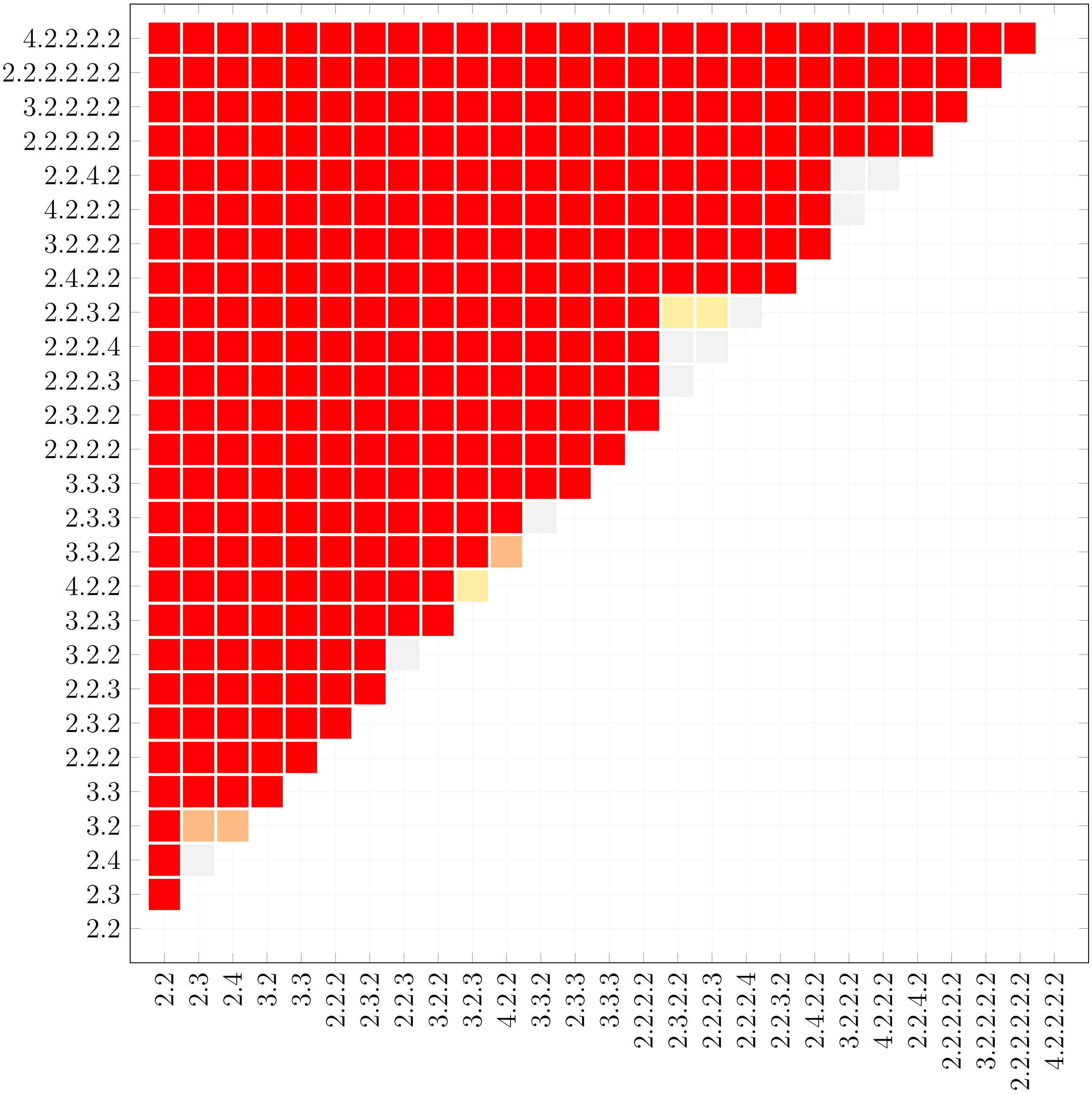


Figure 10: Empirical complexity and length (backward induction).

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 AT1N (average time that the subjects spent at the first node).

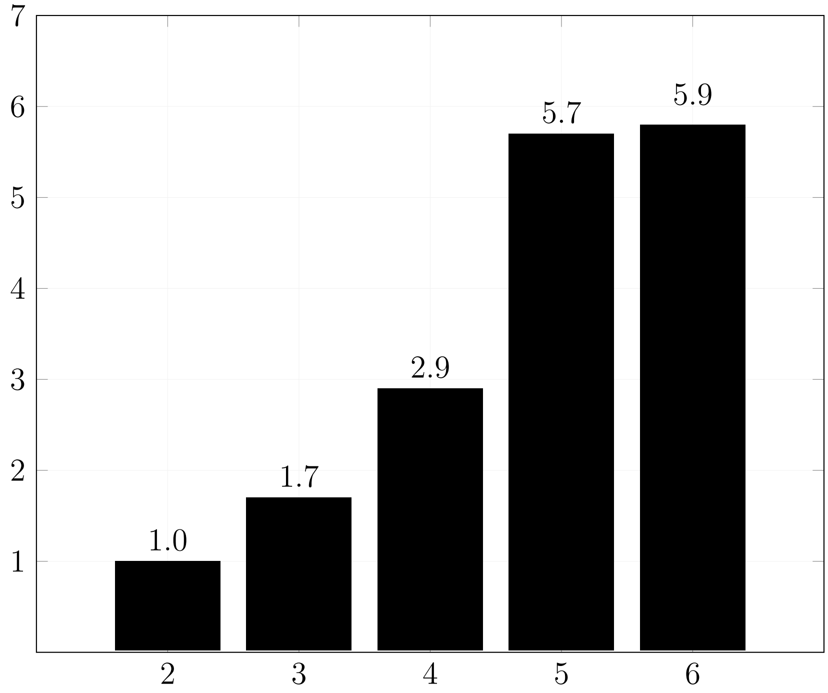


Figure 11: Heat map of empirical complexity (tree construction).

This figure, in a form of a heat map, presents the pairwise comparison of non-trees in accordance with their empirical complexity (average time that the subjects spent at the first node). Non-trees are ranked according to their empirical complexity from the easiest (2.3) to the most difficult (4.2.2.2.2). Colors indicate statistical significance of difference in empirical complexities. Take a non-tree X from the horizontal axis and a non-tree Y from the vertical axis. Red – Y is more complex than X at the 1% significance level, orange – Y is more complex than X at the 5% significance level, yellow – Y is more complex than X at the 10% significance level, and grey – Y is not statistically more complex than X.

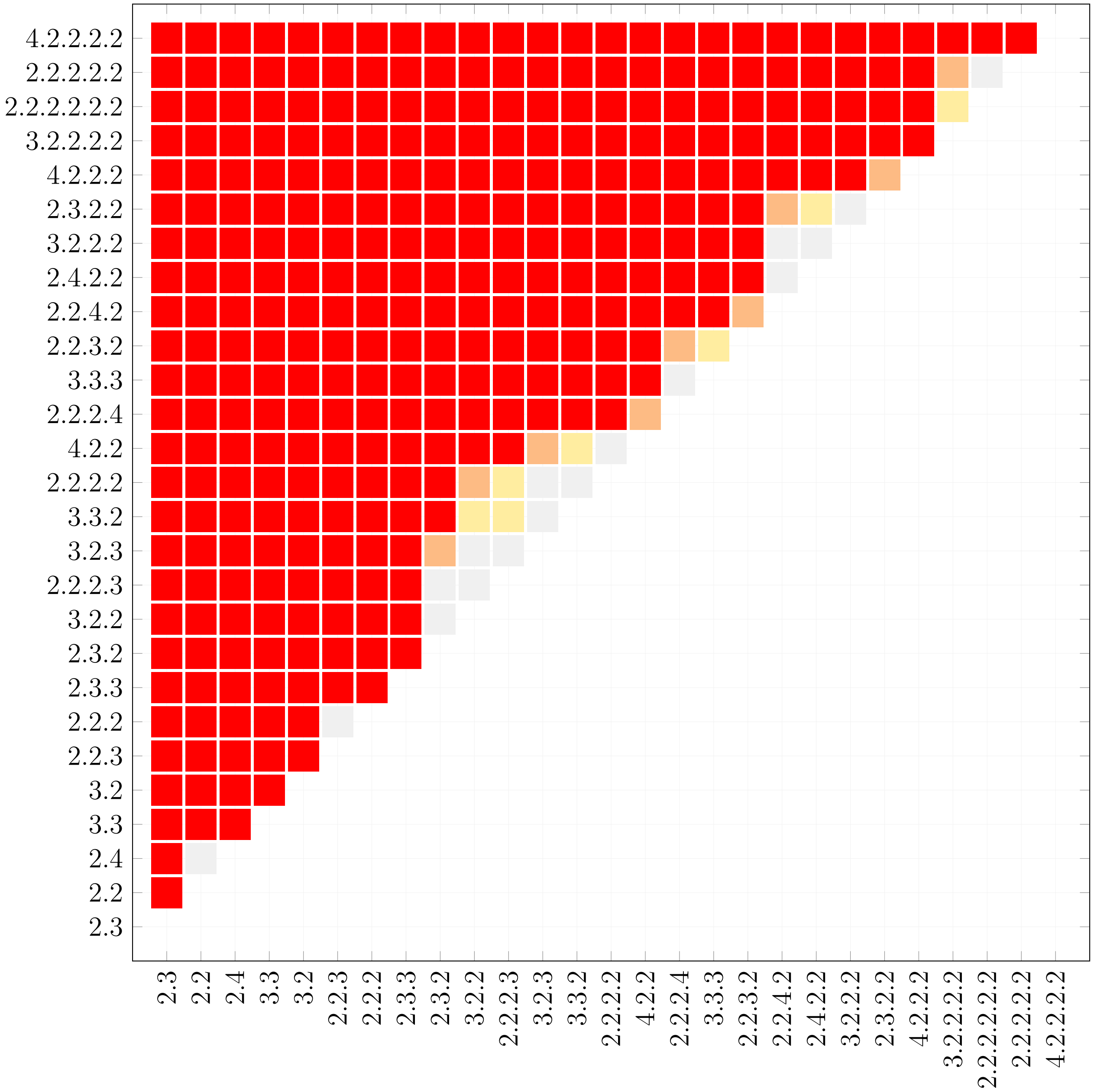
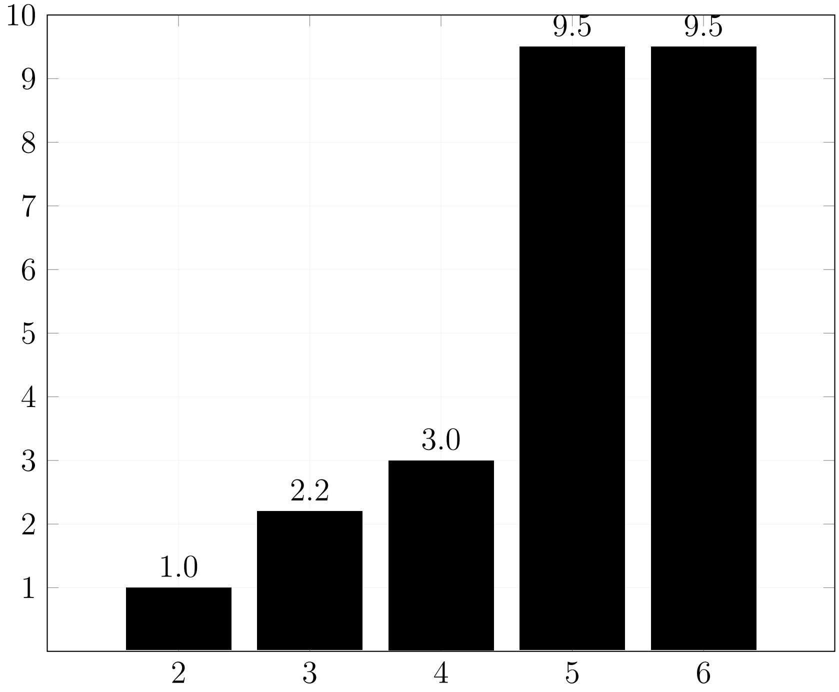


Figure 12: Empirical complexity and length (tree construction).

This figure presents the weighted average of empirical complexity in non-trees with 2, 3, 4, 5, and 6 rounds. Recall that the empirical complexity is measured as AT1N (average time that the subjects spent at the first node).



List of Tables

Table 1: Levels 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		

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

For each tree level, we provide the number of subjects who played, N, and compute the percentage of subjects who did not backward induct, % NOT BI. The data contains 44,113 trees played by 6,677 different players.

level	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%

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

For each non-tree level, we provide the number of backward inducting subjects, N, and compute the percentage of backward inducting subjects who did not construct a tree, % NOT TC. The data contains 28,587 trees played by 4,646 different players.

level	N	% NOT TC
2.2.3	1,248	12.98%
3.3	1,335	14.68%
2.2.2	1,212	15.92%
2.3	1,413	18.12%
2.4	1,343	18.47%
2.3.2	1,178	20.37%
2.2	1,410	21.21%
2.3.3	1,280	21.33%
3.2.3	1,166	24.61%
2.3.2.2	772	34.84%
3.2.2	1,195	35.73%
3.3.2	1,184	36.06%
2.4.2.2	1,103	36.54%
4.2.2	1,173	36.57%
3.2	1,422	37.06%
3.3.3	1,179	38.17%
2.2.2.3	1,023	40.27%
2.2.3.2	1,019	44.65%
2.2.2.2	880	45.00%
2.2.2.4	1,072	46.64%
2.2.4.2	939	49.09%
2.2.2.2.2	690	50.58%
3.2.2.2	875	50.97%
4.2.2.2.2	433	52.66%
4.2.2.2	927	55.66%
2.2.2.2.2.2	472	56.14%
3.2.2.2.2	644	63.98%

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

subject	time spent at the 1st round	time spent at the 3rd round
Ann	15 sec	5 sec
Bob	30 sec	10 sec
Chris	8 sec	12 sec
David	16 sec	24 sec

Table 5: Percentage of backward inducting subjects.

For each tree level, we divide subjects into terciles — L(ow), M(edium), and H(igh) — with respect to a given measure: RT1N (relative time that a subject spent on the first node as a percentage of total time), TT (total time that a subject spent on solving a tree), and T1N (time that a subject spent on the first node).

	RT1N			TT			T1N		
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%

Table 6: Backward induction and TT conditional on RT1N: aggregate analysis.

First, we divide data into terciles — (L)ow, (M)edium, and (H)igh — by RT1N (relative time at the first node). Second, each RT1N-generated tercile, we divide into terciles — (L)ow, (M)edium, and (H)igh — by TT (total time).

		TT		
		L	M	H
RT1N	L	64.67%	53.46%	33.33%
	M	92.53%	89.71%	78.15%
	H	98.39%	97.81%	95.04%

Table 7: Backward induction and TT conditional on RT1N: individual-tree analysis.

		L	M	H			L	M	H
2.2.2	L	93.75%	90.86%	67.55%	2.3.2.2	L	9.6%	3.57%	3.35%
	M	99.55%	98.68%	95.69%		M	80.85%	74.14%	40.32%
	H	100%	99.39%	98.82%		H	98.9%	99.43%	94.94%
2.2.2.2	L	35.87%	8.84%	7.98%	3.2.2	L	88.95%	80.24%	55.56%
	M	95.24%	90.28%	67.82%		M	99.57%	97.79%	96.53%
	H	99.46%	99.45%	93.96%		H	99.35%	98.97%	97.14%
2.2.2.2.2	L	69.23%	34.68%	16.76%	2.2.2.3	L	64.52%	46.43%	21.23%
	M	94.35%	89.82%	56.4%		M	98.41%	97.48%	87.7%
	H	100%	98.85%	91.81%		H	100%	98.89%	95.48%
3.3.2	L	85.55%	82.44%	54.95%	2.2.3	L	93.47%	87.79%	71.78%
	M	99.33%	98.91%	96.22%		M	99.02%	99.39%	93.68%
	H	99.46%	97.92%	95.34%		H	100%	99.49%	97.99%
2.2.2.4	L	64.37%	72.63%	37.02%	2.2.3.2	L	54.7%	43.68%	19.79%
	M	94.79%	96.59%	88.44%		M	97.24%	98.38%	84.38%
	H	98.25%	98.9%	96%		H	100%	98.95%	98.37%
3.3.3	L	84.16%	77.92%	47.85%	2.2.2.2.2.2	L	23.43%	26.4%	13.22%
	M	99.52%	99.44%	95.74%		M	40.46%	37.64%	31.43%
	H	100%	99.46%	96.09%		H	81.71%	85.14%	82.49%
4.2.2.2	L	47.73%	24.47%	10.92%	2.4.2.2	L	69.31%	56.59%	28.89%
	M	93.75%	93.09%	70.86%		M	94.74%	94.41%	85.63%
	H	98.91%	97.75%	95.43%		H	98.94%	100%	96.65%
2.3.3	L	96.22%	90.72%	57.22%	3.2.2.2.2	L	51.46%	25.29%	12.72%
	M	100%	100%	98.38%		M	86.75%	80.22%	60.12%
	H	99.44%	99.46%	98.35%		H	98.24%	95.93%	93.1%
3.2.3	L	90.57%	80.85%	58.2%	4.2.2	L	89.77%	78.17%	48.99%
	M	100%	97.93%	95.98%		M	100%	100%	94.87%
	H	100%	100%	98.27%		H	98.91%	99.07%	97.21%
3.2.2.2	L	34.71%	24.18%	10.53%	2.2.4.2	L	50.85%	29.03%	11.35%
	M	96.45%	94.02%	76.05%		M	98.94%	94.36%	74.05%
	H	98.84%	96.22%	94.86%		H	100%	97.79%	95.74%
2.3.2	L	95.4%	86.96%	63.78%	4.2.2.2.2	L	25.86%	9.94%	2.27%
	M	100%	100%	87.91%		M	61.27%	38.98%	24.86%
	H	99.44%	99.44%	97.71%		H	93.68%	90.23%	84.48%

Table 8: Ranking of trees by potential empirical measures of complexity (backward induction).

For each tree level, we provide the percentage of subjects who did not backward induct (% NOT BI), average total time that the subjects spent solving a tree (ATT), and average time that the subjects spent at the first node (AT1N).

tree	% NOT BI	tree	ATT	tree	AT1N
2.3	2.5%	2.2	9.09	2.2	9.09
2.4	3%	2.3	9.47	2.3	9.47
3.2	3.29%	2.4	9.53	2.4	9.53
2.2	4.4%	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	15.7
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	30.1	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	53.1%	4.2.2.2.2	95.08	4.2.2.2.2	72.27

Table 9: Selecting empirical measure of complexity (backward induction).

We compare each empirical measure of complexity to the objective measure of complexity. Agree: if tree A is objectively more complex than tree B and empirical complexity of A is higher than empirical complexity of B at the 1% level of significance. Disagree: if tree A is objectively more complex than tree B and empirical complexity of B is higher than empirical complexity of A at the 1% level of significance. Undefined: if tree A is objectively more complex than tree B and empirical complexity of A is not different than empirical complexity of B at the 1% level of significance. We analyze three candidates for the empirical measure of complexity: % NOT BI (percentage of subjects who did not backward induct), ATT (average total time that the subjects spent solving a tree), and AT1N (average time that the subjects spent at the first node).

	Agree	Disagree	Undefined
% NOT BI	105	9	22
ATT	132	0	4
AT1N	133	0	3

Table 10: Empirical complexity and width (backward induction).

These panels present the analysis of how tree length and width influence empirical complexity. For each cell in Panel (a), there is an associated cell in Panel (b) with the average empirical complexity. Recall that the empirical complexity is measured as AT1N (average time that the subjects spent at the first node).

Panel (a)

	2 rounds	3 rounds	4 rounds	5 rounds	6 rounds
Min	2.2	2.2.2	2.2.2.2	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	6 rounds
Min	9.09	13.53	21.61	43.33	56.16
Med	9.61	15.50	27.65	48.92	
Max	10.02	18.15	29.55	72.27	

Table 11: Summary statistics for the logit regression (backward induction).

Variable	Mean	St. Dev.	Min	Max
Win	0.78	0.41	0	1
Skills	0.74	0.16	0.02	0.98
Complexity	36.91	19.82	19.05	95.08
Sequence	8.20	7.22	1	27
Number of observations = 35,826				

Table 12: Results from the logit regression (backward induction).

Variable	Logit MLE
Skills	9.892 (0.137)
Complexity	-0.025 (0.001)
Sequence	0.042 (0.003)
Intercept	-5.002 (0.100)
Observations	35,826
Correct Prediction (%): Overall	86.48%
Correct Prediction (%): Backward inducting	88.41%
Correct Prediction (%): Not backward inducting	76.25%

Table 13: Marginal effect from the logit regression (backward induction).

Variable	Margins Estimate
Skills*	19.97%
Complexity*	-6.21%
Sequence	0.52%

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

Table 14: Percentage of tree constructing subjects.

For each non-tree level, we divide subjects into terciles — L(ow), M(edium), and H(igh) — with respect to a given measure: RT1N (relative time that a subject spent on the first node as a percentage of total time), TT (total time that a subject spent on solving a tree), and T1N (time that a subject spent on the first node).

	RT1N			TT			T1N		
tree	L	M	H	L	M	H	L	M	H
2.2.2	69%	88%	96%	93%	86%	73%	83%	89%	80%
2.2.2.2	18%	57%	90%	42%	61%	62%	29%	66%	70%
2.2.2.2.2	24%	43%	81%	41%	52%	55%	32%	50%	66%
3.3.2	37%	64%	92%	75%	59%	57%	56%	64%	72%
2.2.2.4	22%	49%	89%	38%	56%	65%	28%	60%	72%
3.3.3	18%	74%	94%	66%	62%	57%	43%	70%	72%
4.2.2.2	12%	33%	88%	26%	41%	66%	15%	47%	71%
2.3.3	51%	88%	97%	90%	82%	64%	74%	82%	81%
3.2.3	51%	85%	89%	83%	82%	61%	72%	81%	73%
3.2.2.2	15%	43%	90%	34%	50%	63%	24%	49%	74%
2.3.2	57%	87%	95%	86%	82%	72%	71%	84%	84%
2.3.2.2	22%	77%	96%	47%	64%	85%	33%	72%	91%
3.2.2	21%	78%	93%	79%	63%	52%	54%	68%	71%
2.2.2.3	25%	61%	93%	44%	64%	71%	33%	67%	79%
2.2.3	72%	92%	97%	92%	90%	79%	84%	89%	88%
2.2.3.2	17%	55%	94%	51%	54%	61%	28%	65%	74%
2.2.2.2.2.2	15%	32%	85%	17%	52%	61%	15%	45%	71%
2.4.2.2	35%	62%	94%	55%	61%	74%	41%	68%	82%
3.2.2.2.2	10%	25%	73%	25%	35%	49%	14%	34%	60%
4.2.2	22%	75%	93%	69%	62%	59%	47%	66%	78%
2.2.4.2	12%	52%	88%	30%	56%	67%	18%	60%	74%
4.2.2.2.2	20%	39%	83%	37%	45%	60%	27%	43%	72%

Table 15: Tree construction and TT conditional on RT1N: aggregate analysis.

First, we divide data into terciles — (L)ow, (M)edium, and (H)igh — by RT1N (relative time at the first node). Second, each RT1N-generated tercile, we divide into terciles — (L)ow, (M)edium, and (H)igh — by TT (total time).

		TT		
		L	M	H
RT1N	L	45.13%	29.34%	22.05%
	M	74.60%	69.88%	52.60%
	H	96.56%	93.50%	84.45%

Table 16: Tree construction and TT conditional on RT1N: individual-tree analysis.

		L	M	H			L	M	H
2.2.2	L	92.2%	76.12%	41.1%	2.3.2.2	L	20.93%	24.71%	21.84%
	M	98.39%	95.71%	68.55%		M	75%	75.61%	81.4%
	H	100%	97.84%	89.47%		H	94.12%	97.7%	95.35%
2.2.2.2	L	13.83%	20.19%	18.75%	3.2.2	L	51.54%	7.41%	6.02%
	M	73.27%	56.52%	42%		M	96.45%	87.1%	50.38%
	H	98.98%	94.9%	76.29%		H	96.18%	96.32%	87.12%
2.2.2.2.2	L	41.33%	21.52%	9.21%	2.2.2.3	L	16.04%	27.2%	31.82%
	M	51.95%	49.35%	27.63%		M	63.79%	62.73%	56.52%
	H	93.51%	80.26%	70.13%		H	98.21%	89.74%	91.07%
3.3.2	L	68.42%	29.69%	11.85%	2.2.3	L	87.31%	78.42%	52.14%
	M	86.4%	64.58%	40.77%		M	97.9%	94.89%	83.22%
	H	97.67%	96.18%	82.17%		H	98.59%	97.78%	93.33%
2.2.2.4	L	18.97%	18.7%	29.66%	2.2.3.2	L	10.62%	14.04%	26.55%
	M	52.5%	54.55%	40.16%		M	67.29%	62.5%	36.61%
	H	98.21%	92.8%	76.52%		H	98.21%	94.78%	87.61%
3.3.3	L	36.57%	8.53%	7.75%	2.2.2.2.2.2	L	11.32%	13.21%	20.75%
	M	91.2%	83.7%	47.01%		M	19.61%	45.28%	29.41%
	H	97.71%	98.46%	85.61%		H	90.38%	84.91%	79.25%
4.2.2.2	L	5.77%	16%	14.42%	2.4.2.2	L	39.2%	25.21%	39.02%
	M	36.89%	29.91%	32.67%		M	68.33%	62.99%	54.55%
	H	92.08%	87.62%	84.31%		H	99.18%	96.72%	85.48%
2.3.3	L	89.93%	47.59%	16.9%	3.2.2.2.2	L	14.08%	9.72%	7.04%
	M	95.17%	93.96%	73.76%		M	38.89%	20.83%	15.49%
	H	98.59%	97.74%	95.14%		H	84.72%	68.06%	64.79%
3.2.3	L	78.79%	52.8%	21.71%	4.2.2	L	43.08%	6.92%	15.27%
	M	92.31%	93.23%	70.68%		M	92.31%	80.3%	52.71%
	H	98.41%	96.12%	73.64%		H	95.45%	96.09%	88.55%
3.2.2.2	L	14.43%	11.34%	18.56%	2.2.4.2	L	17.48%	6.73%	10.58%
	M	50.98%	47.83%	29%		M	46.6%	57.27%	52.94%
	H	94.74%	90.91%	84.38%		H	93.4%	90.29%	81.73%
2.3.2	L	80.15%	63.71%	27.82%	4.2.2.2.2	L	39.58%	10.2%	10.64%
	M	92.62%	88.73%	78.29%		M	57.14%	33.33%	27.08%
	H	98.43%	97.78%	89.23%		H	93.62%	85.71%	68.75%

Table 17: Ranking of non-trees by potential empirical measures of complexity (tree construction).

For each non-tree level, we provide the percentage of subjects who did not construct a tree (% NOT TC), average total time that the subjects spent solving a non-tree (ATT), and average time that the subjects spent at the first node (AT1N).

non-tree	% NOT BI	non-tree	ATT	non-tree	AT1N
2.2.3	12.98%	2.3	12.86	2.3	12.86
3.3	14.68%	2.2	13.7	2.2	13.7
2.2.2	15.92%	2.4	13.89	2.4	13.89
2.3	18.12%	3.3	15.18	3.3	15.18
2.4	18.47%	3.2	20.11	3.2	20.11
2.3.2	20.37%	2.2.2	33.24	2.2.3	23.99
2.2	21.21%	2.2.3	33.24	2.2.2	23.99
2.3.3	21.33%	2.3.3	43.17	2.3.3	30.14
3.2.3	24.61%	2.3.2	45.62	2.3.2	33.7
2.3.2.2	34.84%	3.2.3	46.85	3.2.2	34.67
3.2.2	35.73%	2.2.2.3	47.42	2.2.2.3	34.92
3.3.2	36.06%	2.2.2.2	48.39	3.2.3	35.79
2.4.2.2	36.54%	3.2.2	48.72	3.3.2	36.34
4.2.2	36.57%	3.3.2	49.93	2.2.2.2	36.64
3.2	37.06%	4.2.2	51.54	4.2.2	38.19
3.3.3	38.17%	2.2.2.4	54.72	2.2.2.4	40.66
2.2.2.3	40.27%	3.3.3	55.82	3.3.3	41.2
2.2.3.2	44.65%	2.2.4.2	61.06	2.2.3.2	43.18
2.2.2.2	45%	2.2.3.2	61.93	2.2.4.2	46.36
2.2.2.4	46.64%	3.2.2.2	62.6	2.4.2.2	47.82
2.2.4.2	49.09%	2.3.2.2	64.89	3.2.2.2	48.28
2.2.2.2.2	50.58%	2.4.2.2	66.28	2.3.2.2	50.47
3.2.2.2	50.97%	4.2.2.2	69.95	4.2.2.2	54.05
4.2.2.2.2	52.66%	3.2.2.2.2	151.95	3.2.2.2.2	104.85
4.2.2.2	55.66%	2.2.2.2.2.2	191.21	2.2.2.2.2.2	144.36
2.2.2.2.2.2	56.14%	2.2.2.2.2	331.31	2.2.2.2.2	184.32
3.2.2.2.2	63.98%	4.2.2.2.2	1251.81	4.2.2.2.2	983.35

Table 18: Selecting empirical measure of complexity (tree construction).

We compare each empirical measure of complexity to the objective measure of complexity. Agree: if non-tree A is objectively more complex than non-tree B and empirical complexity of A is higher than empirical complexity of B at the 1% level of significance. Disagree: if non-tree A is objectively more complex than non-tree B and empirical complexity of B is higher than empirical complexity of A at the 1% level of significance. Undefined: if non-tree A is objectively more complex than non-tree B and empirical complexity of A is not different than empirical complexity of B at the 1% level of significance. We analyze three candidates for the empirical measure of complexity: % NOT TC (percentage of subjects who did not construct a tree), ATT (average total time that the subjects spent solving a non-tree), and AT1N (average time that the subjects spent at the first node).

	Agree	Disagree	Undefined
% NOT TC	88	13	35
ATT	123	3	10
AT1N	124	3	9

Table 19: Empirical complexity and width (tree construction).

These panels present the analysis of how length and width influence empirical complexity. For each cell in Panel (a), there is an associated cell in Panel (b) with the average empirical complexity. Recall that the empirical complexity is measured as AT1N (average time that the subjects spent at the first node).

Panel (a)

	2 rounds	3 rounds	4 rounds	5 rounds	6 rounds
Min	2.2	2.2.2	2.2.2.2	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	6 rounds
Min	13.70	23.99	36.64	184.32	144.36
Med	16.50	30.67	43.62	104.85	
Max	14.53	36.23	47.01	983.35	

Table 20: Summary statistics for the logit regression (tree construction).

Variable	Mean	St. Dev.	Min	Max
Win	0.63	0.48	0.00	1.00
Skills	0.70	0.20	0.00	1.00
Complexity	90.51	174.90	33.24	1251.81
Sequence	9.27	7.45	1	27
Number of observations = 21,664				

Table 21: Results from the logit regression (tree construction).

Variable	Logit MLE
Skills	7.095 (0.114)
Complexity	-0.001 (0.000)
Sequence	0.038 (0.002)
Intercept	-4.568 (0.083)
Observations	21,664
Correct Prediction (%): Overall	77.29%
Correct Prediction (%): Backward inducting	78.81%
Correct Prediction (%): Not backward inducting	73.74%

Table 22: Marginal effect from the logit regression (tree construction).

Variable	Margins Estimate
Skills*	31.69%
Complexity*	-2.65%
Sequence	0.85%

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