Norwegian School of Economics
Bergen, Spring 2019

Children in the Beauty Contest Game

Behaviour and Determinants of Game Performance

Erika Povea and Firuze Citak

Supervisor: Henning Hermes

Master thesis, Economics and Business Administration

Major: Economics

NORWEGIAN SCHOOL OF ECONOMICS

Acknowledgements

We would like to thank our supervisor Dr Henning Hermes for his invaluable guidance and for giving us the opportunity to work on this project.

We dedicate our thesis to our parents Gunay Citak & Serdar Citak and Sonia Gallo & Fernando Povea.

Norwegian School of Economics

Bergen, June 2019

Firuze Citak

Erika Povea

Abstract

This study investigates the behaviour of children aged 8-11 in a beauty contest game with ten repetitions. We observe that choices convergence towards the unique Nash equilibrium over time. Using data on children's elicited beliefs about the actions of their opponents, we find a discrepancy between choices and these beliefs. Besides the general description of behaviour, we apply the model of depth of reasoning, and learning direction theory. In earlier repetitions of the game, choices exhibit on average lower degrees of reasoning compared to the literature on experimental beauty contest games with adult subjects. Moreover, elicited beliefs reveal lower degrees of reasoning than the actual choices. Throughout the game, about half of the children adjust their choices consistent with the predictions of the learning model.

Once we found evidence that children are able to play a beauty contest game, we study potential determinants of the game performance. There is a significant relationship between the understanding ratings of external observers and performance in the game. Further, while cognitive ability is not relevant, empathy skills appear to be a significant determinant. Lastly, we investigate the significance of having stated accurate beliefs and best-responding to them. Most children fail to best respond to their stated beliefs and those who did, win relatively fewer times if their beliefs were inaccurate.

Finally, we complement the analysis with a sample of adults who played the same beauty contest game. The general behaviour of adults are not far from those of children, however, adults converge to the Nash equilibrium earlier in the game. Similar to children, we observe a discrepancy between stated beliefs and choices. In the beginning of the game, adults exhibit, on average, higher degrees of reasoning than children in terms of choices. Around half of the adults show behaviour in support of the learning model, although the percentage is slightly lower than that of children. We found no evidence that empathy is related to game performance as opposed to children, and cognitive ability remains uncorrelated. Adults best respond to their stated beliefs more often than children, such a strategy improves the winning frequency in the game as long as the beliefs are accurate.

Keywords – beauty contest game, belief elicitation

Contents

List of Figures

List of Tables

1 Introduction

Game-theoretic experimental studies investigate the perception of rationality and hence the sense of strategic thinking. The ability of subjects to play a game accurately provides the ground to reveal the behaviour of interest and thus enables the researcher to make sense of the data. As such, this raises the question when the experimental subjects are children: Do children exhibit rational behaviour and are they able to interact strategically? The game-theoretic experimental literature with children was initially based on this concern. The literature grew as the evidence increased that children did not behave entirely random, but showed signs of rationality and strategic behaviour from early childhood to the end of adolescence. The motivation to study applications of game theory with children lies mainly in providing evidence on the concept of rationality in young ages. Our work aims to contribute to the literature on the behaviour of children in interactive strategic settings. More precisely, we examine the behaviour of children in the beauty contest game in order to shed light on whether or how their behaviour resembles that of adults documented in the literature.

To do so, we examine a dataset that has been collected in a study with children between 8-11 years of age who played a board game version of the beauty contest game. The dataset comprises measures of cognitive ability, empathy skills and documents instructor observations in addition to a belief elicitation procedure. Children were instructed to indicate their beliefs regarding the choices of their opponents, once before the game start and once again during the game. Among other assessments, instructors and children evaluated the understanding of the game for every player. The game was played in groups of five players for ten repetitions. The study was designed to simplify the standard game instructions and parameters to adapt it to children, while precisely representing the fundamental elements of a beauty contest game.

The beauty contest game is a dominance solvable game under the assumption of common knowledge of rationality. The players are required to engage in an infinite process of iterated elimination of dominated strategies to reach the unique Nash equilibrium. In experimental applications of the beauty contest game, these theoretical assumptions are rarely satisfied. Thus, the Nash equilibrium is often not reached. Instead, in the case of

repetition, the choices eventually converged towards the equilibrium.

Nagel (1995) proposed another approach, *the model of depth of reasoning*, in order to explain the boundedly rational behaviour observed in the laboratory. The model rests on the idea that subjects eliminate dominated strategies by forming beliefs about others and without the need of infinite iterations. That is, subjects make choices in the game by best responding to the beliefs that they hold about others. The depths of reasoning that players employ essentially mirrors their belief formation process. Research on this topic has shown the importance of the model describing the behaviour in the beauty contest game (Nagel, 1995; Camerer et al., 2004). Consequently, the beauty contest game became a useful tool to study whether or how individuals anticipate the behaviour of others, the concept of bounded rationality and learning through experience in order to explain the eventual convergence to the Nash equilibrium.

The vast majority of the experimental research on beauty contest games considered the actions of players as reflections of their underlying beliefs. On the other hand, evidence from the studies where methods of belief elicitation have been applied cast doubt on this assumption (Costa-Gomes and Weizsäcker, 2008; Sutter et al., 2013; Lahav, 2015). While it remains questionable that elicited beliefs genuinely reflect the underlying beliefs, recent methods developed in an attempt to overcome this limitation (Schotter and Trevino, 2014; Burfurd and Wilkening, 2018). However, directly asking point estimates (as done in our sample) has argued to be the most suitable approach to keep the procedure as simple as possible when the experimental subjects are children (Brocas and Carrillo, 2018b).

We approach our questions of research from several perspectives to the extent of the information that our dataset comprises. The first objective of this study is to find out if children aged 8-11 years are able to play a beauty contest game. We address this question with a descriptive analysis of choices and then, analyse the behaviour under the model of depth of reasoning and learning direction theory. Children show signs of rational behaviour and the distribution of choices converge towards the Nash equilibrium over time. On the other hand, we observe a discrepancy between the choices of children and the average of their stated beliefs. Motivated by this observation, we apply the model of depth of reasoning to the stated beliefs. Comparing the levels of reasoning that children employ and the levels of reasoning that they expect from others, allow us to approach this

mismatch in an alternative way. From the viewpoint of the model, we infer the possibility that the observed differences between actions and stated beliefs can be attributed to players adjusting their choices downwards taking into account their beliefs about others. Moreover, about half of the players during all the repetitions of the game act consistently to the rule proposed by the learning direction theory.

After finding that children exhibit a meaningful behaviour in the beauty contest game, we look into the relevance of certain characteristics to the game performance. First, we explore children's evaluations about the game understanding of each other, together with those of the instructors. The relationship is expected to reveal if players can identify the elements of game understanding, instead of merely recognising the winners of past rounds. However, we did not find enough evidence that supports this behaviour in children. Instead, the instructors proved to anticipate successful performance in the game based on the understanding they perceive from the players.

Next, following the literature that investigates the relationship between cognitive ability and performance in beauty contest games (Burnham et al., 2009; Gill and Prowse, 2016; Brañas-Garza et al., 2012), we explore the relevance of the corresponding measure as well as empathy skills. The beauty contest game emphasises the ability to take the perspective of others in order to perform well. Thus, we consider empathy as a potential determinant and indeed, found that it is related to the winning frequency in the game. Inspired by the preceding findings, we consider that elicited beliefs can provide insights on the game performance. Precisely, we study the role of best responding to one's stated beliefs taking into account the accuracy of the statements. Although the evidence suggests that these variables are negatively related, the relationship with the total wins is not significant unless the stated beliefs are inaccurate. To analyse the relevance of the potential determinants mentioned above, we use ordinary least square regressions. The OLS estimates enables us to discuss the conditional correlations of our variables of interest with the game performance.

Lastly, we investigate to what extent the results we observe in children replicate in a sample with adults. To do so, we use another dataset that contains information of adults playing the same board game version of the beauty contest game. The complementary analysis with the sample of adults provides a way to compare and contrast our findings

with children. The main result from the replication is that despite specific differences regarding the beliefs and actions, both samples converge towards the Nash equilibrium and yield conclusive results in terms of strategic sophistication. Since the literature with children extensively document the impact of age in cognitive and non-cognitive abilities, the differences observed between both samples are not entirely unexpected.

The rest of our study is organised as follows; Section 2 provides a summary of the related literature. Section 3 describes the beauty contest game studies and the data. Section 4 explains the methodology and presents the main results. Section 5 replicates the analysis with the sample of adults and finally, Section 6 concludes.

2 Literature Review

2.1 The Beauty Contest Game

In a regular beauty contest game, N players simultaneously choose a number within the interval [0:100]. Depending on the game design, players may be asked to select only integers or any number within this range. The player whose selected number is closest to the arithmetic average of all chosen numbers multiplied by a parameter p, wins the game. Although most studies are based on the arithmetic average, the median number of all choices is often considered as an alternative. The parameter p is common knowledge and, usually equals to 2/3 or 1/2.[1] In repeated versions of the game, the chosen numbers, the mean (or the median), the winning number (p times the mean or median) and the winner are typically announced at the end of each repetition. In most experimental settings, the winner receives a pre-announced fixed prize. In the case of more than one player sharing the minimum distance to the winning number, the reward splits between the winners. The beauty contest game is a dominance solvable game. The process of iterated elimination of dominated strategies, when rationality is common knowledge, eventually leads to the unique Nash equilibrium. It can be seen at first glance that for a rational player, choosing $100p$ weakly dominates choosing any number above $100p$. If a player knows all other players are rational, she also knows that others will think the same way and no one will

[1] The game is also called "Guess 2/3 of the average game" or "Guessing game".

choose (100p:100]. Consequently, the player knows any number above $100p^2$ is weakly dominated by $100p^2$. If the player knows that all other players knows that all others are rational, then, the player will exclude all choices that are above $100p^3$, and so on. Thus, under the assumption of common knowledge of rationality, (regardless the game is repeated or not) the unique Nash equilibrium when $p < 1$ is all players announcing zero.[2] (Moulin, 1986; Nagel, 1995; Camerer, 2010).

The game took the "beauty contest" name from the passage in Keynes (1936), which commented on the resemblance between the investment behaviour and beauty contest games in newspapers where contestants were required to select the most beautiful faces and the selection closest the to average opinion of the others was rewarded: *"It is not a case of choosing those which, to the best of one's judgement, are really the prettiest, nor even those which average opinion genuinely thinks the prettiest. We have reached the third degree where we devote our intelligences to anticipating what average opinion expects the average opinion to be. And there are some, I believe, who practise the fourth, fifth and higher degrees."* (Keynes, 1936, p.156). In essence, the metaphor describes the theoretical reasoning process that individuals follow in beauty contest games.[3] Nagel et al. (2017) provides an interesting and comprehensive review on the discovery of the beauty contest game. Although the story of the game is composed of a set of intertwined events, the game appeared in the literature for the first time as "Guess the average" in Moulin (1986) and thereafter, the literature continuously evolved with different adaptations of the game. In almost ten years from that, Nagel (1995) was the first to conduct laboratory experiments with beauty contest games, marking a milestone in the research area. The laboratory experiments with beauty contest games revealed a considerable mismatch between human behaviour and theoretical notions.

First and foremost, the Nash equilibrium was rarely selected by the players. Thus, zero was not necessarily the winning number. In the case of repetition, the chosen numbers in most beauty contest games converged towards the Nash equilibrium over time. Such observations led to the introduction of other concepts to explain the observed bounded rationality of individuals within the beauty contest game framework. An alternative

[2] In the case of $p > 1$, the Nash equilibrium requires players announcing 100.

[3] It is worth to mention that in the process Keynes (1936) defines p equals 1 and thus, there exists many Nash equilibrium.

reasoning process to the iterated elimination of dominated strategies involves the idea that players form beliefs about the other players, and base their actions on the best responses to those beliefs. It is assumed that, at the beginning of the game, players do not form beliefs about others or they simply select a random or most salient number over the interval [0:100]. Such behaviour corresponds to a zero-order belief. A player who is one step ahead of those forming zero-order beliefs chooses the best response to it and thus, forms first-order beliefs. The process goes on to second- or even, n^{th} order beliefs, called the levels of strategic sophistication, the process of finite depth of reasoning, the model of iterated best response, the model of depth of reasoning or the level-k model (Nagel, 1995).

The level-k model in experimental beauty contest games (along with other alternative models)[4] has been studied extensively after the seminal paper of Nagel (1995), revealing that players incorporate finite levels of sophistication. As the literature grew, the importance of the model became more evident and it turned out to be an essential tool for examining several different aspects of the beauty contest game. Duffy and Nagel (1997); Kocher and Sutter (2004); Coricelli and Nagel (2009); Müller and Schwieren (2011); Agranov et al. (2013); Sbriglia (2008); Lahav (2015) are a few notable examples of work in this domain.

Besides the level-k model, some researchers have also examined the learning process of players in beauty contest games in order to explain the eventual convergence to the Nash equilibrium. Nagel (1995) and Duffy and Nagel (1997) applied the learning direction theory to their experimental data where their point of departure was the absence of increasing levels of reasoning, especially in the first few repetitions of the beauty contest game. Since then, several other learning models and methods have been suggested and tested in the experimental beauty contest games to explain the observed convergence behaviour (Stahl, 1996; Camerer and Ho, 1998; Camerer, 2010; Kocher et al., 2014). This behaviour has also been associated with comprehension of the game rules. Duffy and Nagel (1997) compared the behaviour in different adaptations of the beauty contest game, using the mean and the median (and the maximum) of all choices as the game rule. They found greater decrease rates towards zero in the initial rounds of in their median game compared to the mean game. This finding suggests that using median instead of the mean as the game rule facilitates the reasoning process of players. Using different presentation

[4]Camerer et al. (2004); Bosch-Domènech et al. (2010); Breitmoser (2012).

structures of the beauty contest game, Chou et al. (2009) found that, overall, when the game instructions are framed in a simpler way, subjects play (weakly) dominant strategies.

Another branch on beauty contest games adds an appealing dimension to the literature by investigating the relationship between players' cognitive ability and their reasoning process. Burnham et al. (2009) found that higher cognitive ability is linked to choices approaching the Nash equilibrium. Gill and Prowse (2016) confirmed the finding using the Raven test. Brañas-Garza et al. (2012) measure the cognitive ability of experimental subjects both with the Raven test and CRT (Cognitive Reflection Test). They conclude that while the Raven test scores are not associated with the performance or reasoning in the beauty contest, subjects with higher CRT test scores are more inclined to play towards Nash equilibrium. Lastly, Coricelli and Nagel (2009) found significant links between beauty contest game players engaging in different levels of reasoning and their respective brain activity, extending the research area to neuroeconomics.

2.2 Belief Elicitation

In guessing games, the chances of winning depend not only on one's actions but also on the decisions of others. Therefore, there are incentives to form beliefs about opponents' behaviour. However, there are limitations to learn about beliefs when we just study the choices of players in the beauty contest game. The reason is that the structure of this game has a binary classification for actions; one either wins or not. Hence, if we assume players' actions are driven by their beliefs we would not be able to distinguish those subjects with thoughts that are similar to the ones of the winner. Every other action, and thus beliefs that are different from the winner would be treated as equals. Nor will we know if the players end up following the beliefs they formed about others.

Subjects' underlying beliefs are latent variables, but laboratory settings can be useful to transform them into observables through elicitation methods. Belief elicitation is a method in which an experimenter asks the subjects to report their beliefs. However, the accuracy of the statements remains debatable and for this purpose, there are many techniques to incentivise players to state their true beliefs. The method is becoming popular since it provides the opportunity to extract valuable information on unobservables.

Nevertheless, it remains uncertain if elicited beliefs are equivalent to true underlying beliefs. On this regard, Costa-Gomes and Weizsäcker (2008) examined data on normal-form games and proposed a model of behaviour allowing stated beliefs to differ from underlying beliefs, and both of them, from actions. The datasets of each game led to different results and the authors concluded that subjects could state beliefs which are different from their true latent beliefs, as well as someone, can end up making a choice different from the planned. The study of Costa-Gomes and Weizsäcker (2008) has served as a benchmark to further develop models of behaviour relaxing the assumption that choices strongly depend on beliefs. Since then, the evidence that supports the robustness of the results about the discrepancy between beliefs and actions has gradually increased (Sutter et al., 2013).

The literature is far scarce on belief elicitation in beauty contest games. To the best of our knowledge, Lahav (2015) is the only study to conduct a belief elicitation procedure in beauty contest experiments. The author carried out the experiment with two treatments and a control without belief elicitation. The first treatment required players to indicate the frequencies at which certain ranges of numbers will be selected. This question only appeared once at the end of the game. In the second treatment, belief elicitation was applied after every round and the procedure allowed subjects to change their previous choices in the game. Comparing the results of both treatments and the control group, the author found that belief elicitation did not alter the choices in the game. Moreover, the assumption of consistency between actions and stated beliefs was rejected. While Lahav (2015) claimed that actions do not reflect a player's beliefs (at least the ones reported), Schotter and Trevino (2014) discuss several studies where elicited beliefs guided actions in different experimental games (Nyarko and Schotter, 2002; Danz et al., 2012; Hyndman et al., 2013; Manski and Neri, 2013).

2.3 Children in Experimental Economics

Experimental economic studies with children and adolescents have seen an increase in popularity over the past fifteen years as a consequence of providing valuable evidence on economic behaviour. The literature on this topic has evolved around various aspects of behaviour such as competitiveness, rationality, time, risk, and social preferences all of which are well documented experimentally with adult participants. The age range of

the experimental subjects in studies with children varies from early childhood to the end of adolescence. Fehr et al. (2008); Moreira et al. (2010); Brocas and Carrillo (2018a,b); Khadjavi and Nicklisch (2018); Hermes et al. (2019) are a few notable examples considering early childhood (from 3 years of age) in terms of diverse aspects of behaviour. Consequently, such studies bring along different caveats in experimental designs for children, for instance, simplified games and detailed explanation procedures. The research also provides a basis for examining the impact of age on the development of economic behaviour (Sutter et al., 2019).

One of the first studies in this area, Harbaugh et al. (2001) ran an experiment with 7 and 11-year-old subjects in order to investigate whether their choices are consistent with the generalised assumption of revealed preferences. They found evidence of rational behaviour even for their youngest age group, with an increasing trend in age. More recently, concepts such as strategic thinking, elimination of dominated strategies and forming beliefs that constitute the basis of our work have been studied with children where the age group we examine (and even younger age groups) has found extensive coverage (Brocas and Carrillo, 2018a,b; Apesteguia et al., 2018; Barash et al., 2019; Czermak et al., 2016; Brosig-Koch et al., 2015).

Although the findings in the literature favour signs of rationality, strategic behaviour in interactive games, and anticipating opponents' decisions in children, most of the conclusions agree that such features develop with age. Namely, our behaviour of research interest is present even in the youngest age groups, but it becomes more noticeable through adolescence. This observation also depends on the simplicity or complexity of the tasks used in the laboratory with children. Regarding beauty contest games, as far as we know, there is no specific information in the literature of children. One reason for this can be that the beauty contest is not the most straightforward game to comprehend, however, possible to simplify. To that end, we particularly investigate whether children can play and if so, to what extent a beauty contest game.

3 Data

3.1 Children Dataset

The data we analyse in our thesis work comes from the study designed and conducted by Dr Henning Hermes[5] between March and April 2016 in Germany. The study was conducted with six cohorts of third and fifth grade students drawn from three different schools. Parental consent forms for data use were sent to the teachers following the approval of participation from schools. The final sample consisted of 114 children where 64 of them were third graders and 63 of them were female. The ages among participants ranged from 8 to 11 and the mean age was 10 years.

The study was designed to be completed on one school day. The days were organised in the same way for each of the six cohorts. Trained research assistants, called interviewers from now on, guided and observed the children throughout the study. During the first school hour, children received information about the day plan and the rewards they could win at the end of the day by earning coins in the game. Then, children were guided through a workbook which contained various tests and questionnaires. All instructions were read out loud by the interviewers. The workbooks were distributed randomly in the classroom, each with a number on it, which later determined their groups in the game.

In the following school hours, groups of five children were taken out of their classroom to another room to play the game. All groups were formed by selecting children randomly from the same cohort.[6] During the last school hour, children were given the opportunity to exchange the coins they won for toys. In the end, 23 groups of five children played the game. Among those, one child did not provide the parental consent for data use, leaving us with a final sample of 114 participants.

[5] NHH Norwegian School of Economics, henning.hermes@nhh.no

[6] One session included a child from a different cohort due to time reasons. For another session, two children from the second grade were recruited to fulfil the number of participants. Results remain the same when those observations were excluded.

3.1.1 The Goblin Game

Consider the following beauty contest game:

Five players simultaneously choose an integer in the closed interval [0,100]. The median number of all five chosen numbers is determined and multiplied by a parameter p which equals 1/2*, and it is common knowledge. The player whose number is closest to p times the median (i.e. half of the median) wins the round. The game is repeated for ten rounds by the same group of five players. At the end of each round, the chosen numbers, the median number and the winner are announced. The winner of each round earns a coin. If there is a tie, all winners receive a coin.*

The Goblin Game represents the beauty contest game described above in the form of a board game shown in Figure 3.1. In the illustration, the outer circle in black and white provides all the possible integers that players can choose arranged in ascending order from 0 to 100. The inner circle in green displays the numbers from 0 to 100 but in steps of 1/2 to facilitate the estimations and point out the winner.[7] The treasure chest in the centre represents the golden coins children can win at the end of each round. The goblin stands at its initial position zero and waits for the game to start. The game was organised in groups of five children playing on a table where the seating positions were arranged according to five different colours: yellow, blue, orange, white and grey. Each child was randomly assigned to one of these colours which determined their seating positions and represented them throughout the game. Children received a pawn of their colour to place simultaneously on the outer circle of the board to indicate their chosen numbers in each round.

In each session, five interviewers participated in guiding and observing children during the game. Before the game starts, each of the five children sat on separate small tables in the classroom and received a one-to-one standardised explanation of the game rules from an interviewer. During the explanation, interviewers used a small replica of the game materials, i.e. the board, pawns, coins as well as the written instructions. Before explaining the game rules, interviewers reminded the children that the more coins they earn, the more toy options they would have in the end. The idea was to motivate children

[7] Even if they will never be used, the half steps in green cover the circle and go up to 100 to avoid giving hints to participants.

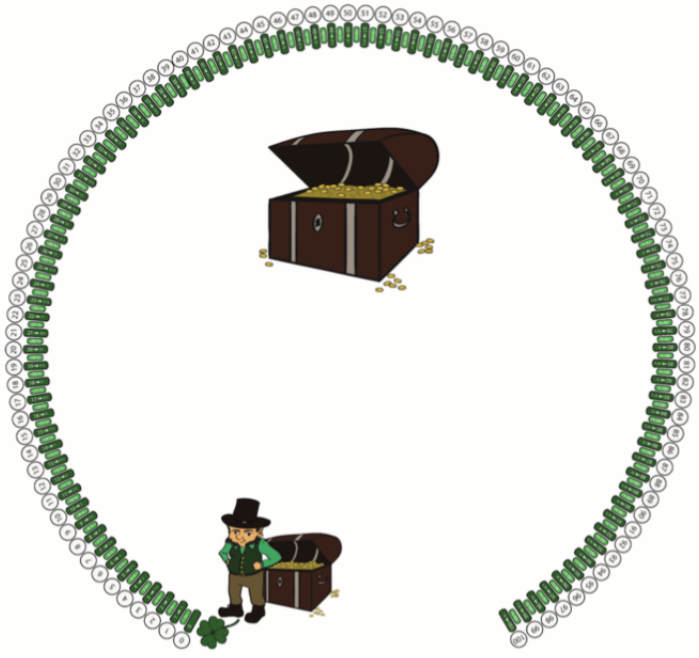

Figure 3.1: The Goblin Game Board

to do their best to earn coins and make them understand that the game is not "pure luck" but they can "do something" to win.

The rules of the game were presented to children like a story: The goblin keeps golden coins in the forest and he will tell the location of the coins to the player closest to him at the end of each round. The goblin walks until the middle (the third) player along the road. The goblin is bewitched, so he has to take half step backwards for each step he takes forwards. The rules, at the same time, were organised into the following five steps:

1. All players secretly write down a number between 0 and 100 on an individual sheet. To indicate that they have made their decision, they cover the written number with their pawn.

2. All players simultaneously place their pawn on the board, of course, on the number they wrote down on their sheet (on the black-and-white numbers).

3. The goblin starts from zero and walks up to the third player. The goblin uses the green, inner circle of numbers.

4. Having reached the third (middle) player, he jumps back by the half of the number the third player was standing on.

5. The player who is now closest to the goblin wins a golden coin. If several players are equally far away from the goblin, they all receive a coin.

Children then had the opportunity to ask questions in private if something was not clear. After explaining the game rules, the interviewer asked the child to describe the five steps written above, back to her. For each of the steps, the interviewer ranked the child's understanding. At the end of the one-to-one explanation, the interviewer asked the child to draw an E on her forehead and documented the result. All children at the end of the explanation received a golden coin.

Once all children were done with receiving instructions in private, they sat on the table centred in the classroom to play the Goblin Game. The main interviewer of the session joined the table to guide children through all ten rounds of the game while the other interviewers observed them. Children first answered some questions in their workbook. Then, the rules were repeated by the main interviewer in five steps. During the fourth step, the main interviewer yawned visibly, and the other interviewers documented if the yawning was contagious among the children.

During the game, the main interviewer moved the goblin according to the numbers chosen by children and gave a golden coin to the winner(s) at the end of each round. The interviewer read out loud the numbers chosen by all five children in each of the rounds and another interviewer documented the choices. In addition, interviewers collected the sheets where children wrote down their numbers. Information corresponding to coins earned and winner(s) colours was also collected along the round. The game paused only for a short while at the end of the fourth round for children to answer some questions in their workbooks. Lastly, all children received an extra golden coin for their participation.

3.1.2 Belief Elicitation and Questionnaires

The information provided by individual children questionnaires and interviewers observations reported in booklets constitutes a significant part of the data in this study. Moreover, teachers filled a survey where they rated math, German and empathy skills of every child on a four-point scale.

Children Questionnaires

Children answered several questions throughout their entire participation. First, they filled a workbook before leaving their classroom to play the Goblin Game. Among other measures, it contained a test of intelligence and two self-rated measures of empathy. Precisely, the Raven's Progressive Matrix Test was used to measure fluid IQ and empathy skills were measured by the FEAS scale[8] and the Interpersonal Reactivity Index (IRI). The Raven test captures the ability to solve logical problems that can not be performed automatically or using previous knowledge. Being a non-verbal test based on images, it also measures the capacity of abstraction; to form new concepts, recognise relationships between patterns, make inferences, and solve problems. FEAS scale contains six small stories designed to capture empathy and the resulting social behaviour and, IRI assesses distinct components of empathy (Gilet et al., 2013).

Once children were seated on the game table, they filled specific items in their questionnaires. There were questions designed to elicit the opinions of children regarding the other players and themselves. Beliefs were elicited twice by using the same questions during the game, once before round 1, and again before round 5. Each of the questionnaire items had an illustration of the game table, seating positions and the respective colours of the children in pawn figures (Figure 3.2). The questions were written on top and there were empty boxes next to each of the pawns for children to write down their answers.

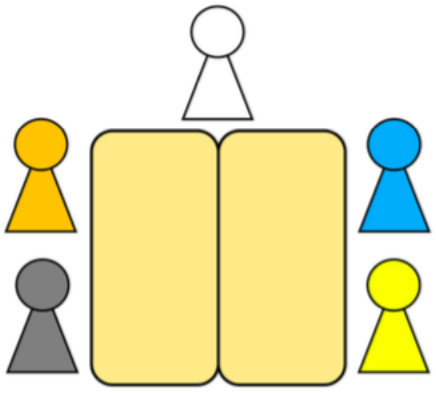

Figure 3.2: Illustration of The Goblin Game Table

The first item of the questionnaire asked children to rate the understanding of each player, including themselves, according to their personal assessment in a four-point scale: *"This child understood the game well... "*. Then, children were asked to select only one of the

[8]Initials in German for "Fragebogen zur Erfassung von Empathie und Angemessenem Sozialem Verhalten" (Meindl, 1998).

players that they think will win the game at the end: *"This child will win the game at the end..."*. The remaining question required children to specify the numbers that they think each of the other players (and themselves) will choose in the next round: *"This child will write down the following number in the first (fifth) round..."*

Interviewer Booklets

In addition to the self-rated measures of empathy, interviewers conducted two behavioural tasks (E on the forehead and contagious yawning) described in Section 3.1.1 to evaluate empathy skills. Interviewer booklets consisted of the documentation of stepwise comprehension ratings, E on the forehead and contagious yawning tasks. Besides, they ranked each child's understanding of the game based on their general impression both before and during the game where the questions were in the form of Figure 3.2. One interviewer per child documented the ratings of stepwise comprehension and the E on the forehead task in the one-to-one explanation session. Conversely, the understanding ratings and the yawning task were documented simultaneously by the impressions of four interviewers present during the session.

3.2 Adults Dataset

The second set of data was collected in the study where adult participants played the Goblin Game. The study was conducted in July 2018 and the participants were students of the Johannes Gutenberg University Mainz. The final sample consisted of 120 subjects where 72 of them were female, and the mean age was 23 years. Although the study was similar to the one conducted with children, it differed in several ways. First of all, participants played a regular beauty contest game ($p = 2/3$). Then, the subjects answered questionnaires and finally proceeded to the Goblin Game.

The questionnaires consisted of the Raven test for cognitive ability and IRI to measure empathy. Having completed the tests, participants were randomly assigned colours (yellow, blue, orange, white or grey) to represent them during the Goblin Game and to determine their seating positions. The game rules were presented to the participants in a similar way as in the study with children. Subjects were told the story of the goblin and, the

rules further explained in five steps.[9]

Naturally, the explanation procedure was standard and not as detailed compared to the study with children. Five subjects played the Goblin Game for ten rounds. As the total number of participants was 120, we end up with 24 sessions. After each round, the chosen numbers, the median number and the winner were announced. A researcher accompanied the players during the game to move the goblin and to make the announcements. The winner (or winners) of each round received a symbolic golden coin. At the very end of the game, one of the ten rounds was randomly selected. The winner of the randomly drawn round received 20 EUR. If there was a tie, the prize was shared between the winners.

The belief elicitation procedure was similar to the study carried out with children, except that adults only answered questions about choices once before the first round and again before the fifth round. Players received the question on paper: *"This participant will write down the following number in the first (fifth) round ..."* The question included the illustration of the game table (Figure 3.2) and answer boxes corresponding to each of the other players.[10] The major difference of the belief elicitation procedure in adults was the reward received for the accuracy of beliefs. That is, the distances between their stated beliefs and the corresponding numbers that others chose. Participants received an additional payment of 1 or 0.5 EUR if their second best distance[11] was smaller than five or ten, respectively. Lastly, after the payout procedure, researchers conducted the "E on the forehead" task with the participants and documented the results.

4 Results

In this section, we first present the descriptive statistics of children's behaviour throughout the Goblin Game. In the first two subsections, we consider the first round behaviour of children in detail and then, focus on further repetitions of the game. We demonstrate that children on average share many similarities in terms of choices with the behaviour of adults observed in standard versions of the beauty contest game. What is not apparent

[9] 1. Secretly write down an integer between 0 and 100. 2. Simultaneously place your pawn to the corresponding number on the game board. 3. The goblin runs to the third player. 4. The goblin jumps half-way back. 5. The player closest to the goblin wins the round.

[10] Adults did not provide answers concerning themselves in the belief elicitation procedure.

[11] The researchers awarded the second best distances to avoid collusion among the participants.

are the mental processes underneath and how they interact to achieve these outcomes. For that reason, we include an analysis of strategic behaviour using the model of depth of reasoning and the learning direction theory. Results on actions and stated beliefs provide the foundations to explore links between game performance and various measures; such as understanding of the game, cognitive ability, and empathy. Finally, we investigate how children react to their stated beliefs and evaluate them against game performance.

4.1 Descriptive Statistics

4.1.1 Behaviour in Round 1

In the first round, the median of the chosen numbers across all sessions is 28, the mean is 33, and the standard deviation is 20.6. Consequently, the distribution of choices is widely spread over all intervals. No players choose zero in the first round and only 6% of choices are below 10 (Figure 4.1). The interval of dominated strategies (choices above 50) has the highest concentration of observations in the first round (18%) and 5% of the players choose exactly 50. The second and third highest bars in Figure 4.2 indicate that 28% of children's choices lie between numbers 15 and 25. The observed behaviour in the first round of the Goblin Game is consistent with previous beauty contest game experiments where first round choices are in general far away from the Nash equilibrium. For instance, our results resembles the frequencies of first round choices in Nagel (1995) to a large extent. Moreover, Camerer et al. (2004) ran several experiments using the beauty contest game and found that the average was often within the interval 25-40 with a large standard deviation of around 20.

Results in the initial round may not be sufficient to draw conclusions about the behaviour of the players, particularly since they do not have information about their opponents' behaviour. The first round is fundamental for individuals to form expectations about the other players and to adjust their behaviour accordingly. Although we cannot consider results from round 1 as conclusive, we will use them as a benchmark to compare the development of the behaviour in the remaining rounds.

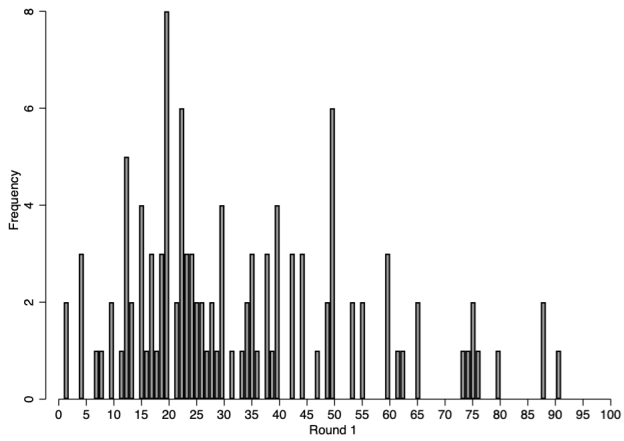

Figure 4.1: Chosen Numbers in Round 1

4.1.2 Behaviour in Rounds 2 to 10

Figure 4.2 shows the distribution of choices with combined histograms for each of the ten rounds of the Goblin Game.[12] Chosen numbers are grouped in intervals of five except for the last, which contains all dominated strategies (50:100]. The y-axis shows the frequencies of the selected numbers. The rounds that register the highest rate of numbers below five are 6-8. There is a noticeable drop in choices larger than 50 after round 1, which suggests that most individuals identified dominated strategies after playing the first round of the game. The spread nature of the choices in the initial rounds fades as the game repeats.

Despite the undefined behaviour mostly observed in the first round, children's choices converge towards the Nash equilibrium in a similar way as adults do in standard versions of the beauty contest game. The share of children choosing zero changed from no players in the first round to 40% in round 10.[13] Not only the selection of numbers approaches the theoretical equilibrium over rounds, but the dispersion of choices also narrows, reaching a standard deviation of 2.5 in the final round.

Figure 4.3 plots the transitions[14] of chosen numbers between consecutive rounds (from round t to round $t+1$) over the game. If a player chooses a lower number in the next

[12] A more detailed visualisation of the chosen numbers for each round of the Goblin Game can be found in Appendix A.

[13] From round 1 to round 3, no players chose zero but in round 4 three of them played the Nash equilibrium. The number of subjects with this choice increased to eight in round 5, to six in round 6 and then to 20 in the seventh round. The frequency of participants choosing zero continued to increase from 27 (round 8), to 35 (round 9) and finally up to 46 in the last round.

[14] Nagel (1995), Kocher and Sutter (2004).

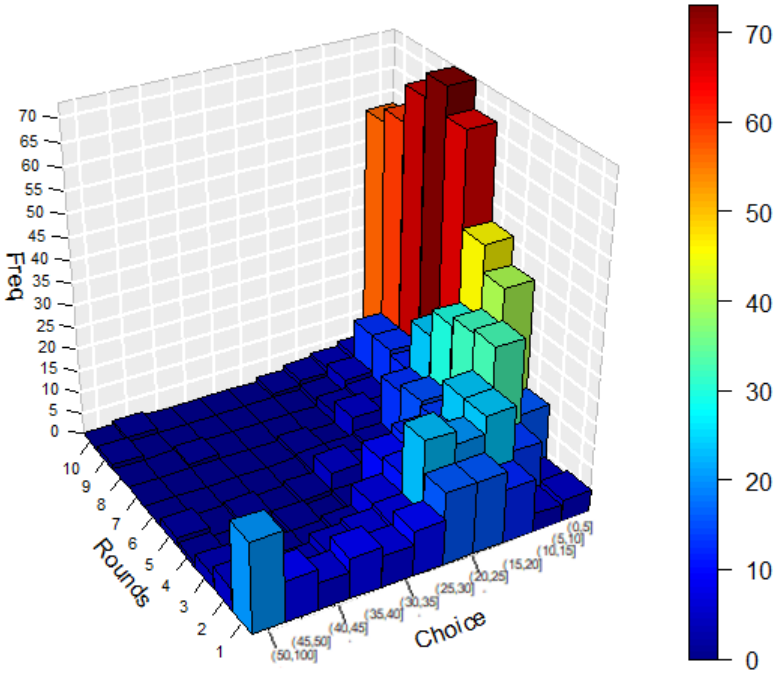

Figure 4.2: Chosen Numbers in the Goblin Game

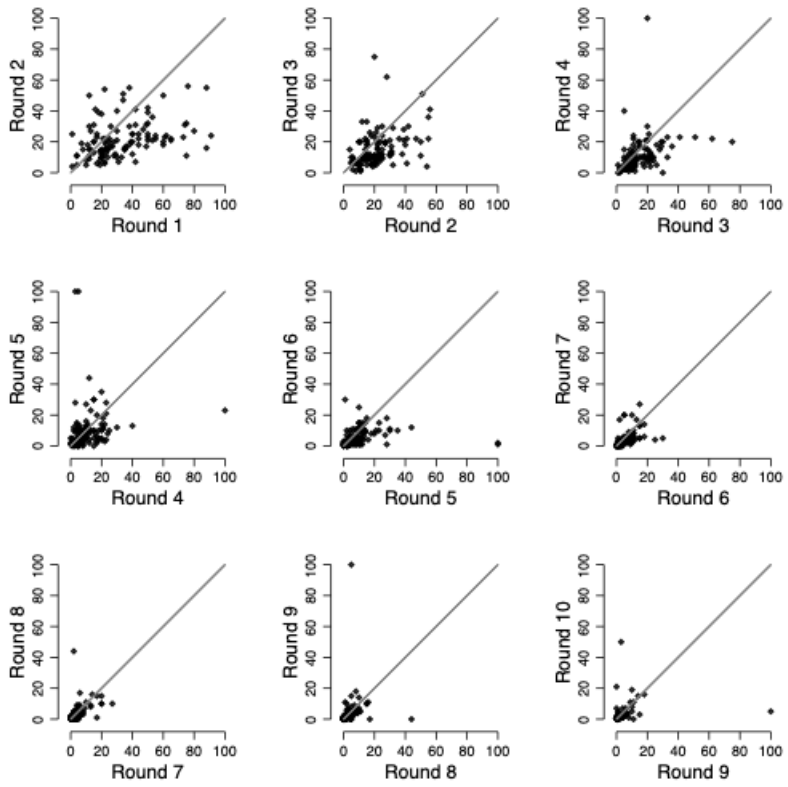

Figure 4.3: Transitions of Chosen Numbers from Round t to Round $t+1$

round compared to the current round, the observation would lie below the 45-degree line. Indeed, the choices of players decrease from any round t to the next round $t+1$ (signtest, $p \approx 0.00$). The greatest amount of observations (85 out of 114) that lie under the diagonal are between rounds 2 and 3. Even though in the last rounds there are still choices higher than zero up to 20, and outliers; the observations that are widely outspread in the beginning of the game approach, on average, the Nash equilibrium.

Table 4.1: Medians and Means of Chosen Numbers Over Rounds

	Median	1-Median(t)/Median(t-1)	Mean	1-Mean(t)/Mean(t-1)
Number Round 1	28		33.46	
Number Round 2	20	0.29	21.73	0.35
Number Round 3	11.5	0.43	14.96	0.31
Number Round 4	8	0.30	10.68	0.29
Number Round 5	6.5	0.19	10.04	0.06
Number Round 6	3	0.54	5.42	0.46
Number Round 7	2.5	0.17	3.89	0.28
Number Round 8	2	0.20	3.36	0.14
Number Round 9	1	0.50	3.33	0.01
Number Round 10	1	0.00	2.51	0.25

Table 4.1 shows the medians and means of all chosen numbers per round. The decrease of both, the median and the mean, between any consecutive rounds confirms that the numbers chosen by children approached the Nash equilibrium as the game repeated. Mean values are significantly higher than the medians over the ten rounds (signtest, $p \approx 0.00$), capturing the fact that the mean is a more sensitive parameter to positive outliers observed in Figure 4.3. The second and fourth columns of the table show the percentage changes in medians and means from round $t-1$ to round t (Nagel, 1995). The decrease rates of medians and means, on the other hand, are not significantly different from each other (signtest, $p > 0.50$), meaning that both measures converge at a similar rate towards the Nash equilibrium.

4.1.3 Stated Beliefs

In this section, we provide a summary of the elicited beliefs and compare them to the observed behaviour in the game. As described in Section 3.1.2, children reported beliefs

about every other player in the session, including themselves.[15] The belief elicitation procedure was conducted twice, once before round 1 and once again before round 5. Each question about beliefs, which includes understanding of the game and the expected choices, was designed to provide five observations per child, i.e. 570 observations (5 x 114 children), since all children were asked to indicate an answer for themselves and the other four children in the session. The question about the final winner required children to specify only one participant; thus we get one observation per child. Though, not every child provided a valid answer to all of the questions which left us with missing observations. For this reason, the sample size varies throughout the analysis depending on the number of incomplete answers in the questionnaires.

We begin with the question concerning the winner of the game: *"This child will win the game at the end..."*. Before round 1, 47% of the players indicated themselves as the winner of the game. After the fourth round, this percentage dropped to 30% suggesting that they became less confident about themselves. In contrast, there is almost no variation in children's answers about themselves to the question: *"This child understood the game well..."*. Children used ratings on a four point scale where 1 means "not at all" and 4 is "yes fully". More than 93% of children expressed that they fully understood the game and this did not change —except for nine players— from round 1 to round 5. Even if children became less confident about winning, their perception of the game understanding was particularly high and remained the same. Thus, we cannot consider the self-regarding measure of understanding as informative to explain behaviour over time. We analyse children's other-regarding ratings of understanding in Section 4.3.2.

Table 4.2 shows the pairwise correlations between the number of times each child believes a player will win at the end of the game and the corresponding winning frequency aggregated in different rounds: Coins 1 to 10 (number of coins won in total over the ten rounds), Coins 1 to 4 (number of coins won from round 1 to round 4) and Coins 5 to 10 (number of coins won from round 5 to round 10). The table also presents the relevance of winning in different parts of the game with regards to the final winning frequency (1-10). Rows (1), (2) and (3) of the correlation table show that the number of coins won during the last

[15] In the rest of our analysis, we exclude children's self-regarding beliefs about the choices and focus solely on their beliefs about the choices of others. The analysis that led us to this consideration can be found in Appendix B. The expressions *elicited beliefs* or *stated beliefs* refer to other-regarding beliefs hereafter.

six rounds of the game has a stronger correlation with the final winning frequency (Coins 1 to 10) than the moderate relationship revealed with the first four rounds of the game. Moreover, the low and insignificant correlation (0.11) demonstrates that the winning frequencies in the first four rounds and the last six rounds are not related. This means that those children who won more often at the beginning of the game were not necessarily the ones who accumulated more coins at the end. Both correlations are significant at a 1% level.

Table 4.2: Pairwise correlations – Winning Frequency and Beliefs

	(1)	(2)	(3)	(4)
(1) Coins 1 to 10				
(2) Coins 1 to 4	0.54***			
(3) Coins 5 to 10	0.90***	0.11		
(4) Winner Belief R1	-0.01	0.12	-0.07	
(5) Winner Belief R5	0.35***	0.71***	0.05	0.23**

Note: The stars indicate significance levels: *p $<$ 0.10, **p $<$ 0.05, ***p $<$ 0.01.

The variables in rows (4) and (5) of Table 4.2 summarise children's answers to which player they think will win the game in the end. Thus we define a variable that compiles the "votes" every child received from their peers which also includes cases where players voted for themselves. By doing so, we can correlate the general opinion of children with the number of coins and check how their ability to identify winners change from round 1 —where we assume chances of having accurate predictions are low— to round 5.

The insignificant relationships between the winner belief in round 1 and the number of coins accumulated at the end of the game indicates that children could not guess the winner accurately at the beginning. Having played four rounds, the correlation between children's beliefs about the winner (winner belief round 5) and the actual winning frequency in the past (Coins 1 to 4) becomes strong and significant (0.71). However, we do not observe such a relationship between the winner belief in round 5 and the future winning frequencies (Coins 5 to 10). This suggests that information on past winners influenced most votes and the majority of children in round 5 voted for the player who won more often during the first four rounds. In other words, children are aware of who is winning i.e., they are paying attention to the development of the game but they are not able to anticipate the prospective winner.

Moving on, the question *"This child will write down the following number in the next round..."* requires children to state their beliefs about the number each of the players will choose. In order to summarise the beliefs about the opponents in a comparable manner to the actual choices in the game, we calculate the arithmetic mean of each child's stated beliefs. The mean value takes into account the effect of each observation unlike the median. We graphically represent what each player thinks on average about the behaviour of others against their actual choices (Figure 4.4). If children played a number that corresponds to their average beliefs about the opponents, the matched observations would be centred around the 45-degree line. This is not the case for neither the first nor the fifth round. Nevertheless, the fact that most observations lie below the diagonal indicates that children often played lower numbers than the average beliefs about their opponents. In round 1, 74% of the observations lie below the diagonal, and in round 5, the percentage increased to 88%. Thus, the behaviour of choosing lower numbers than their beliefs persisted and became even larger as the game repeated.

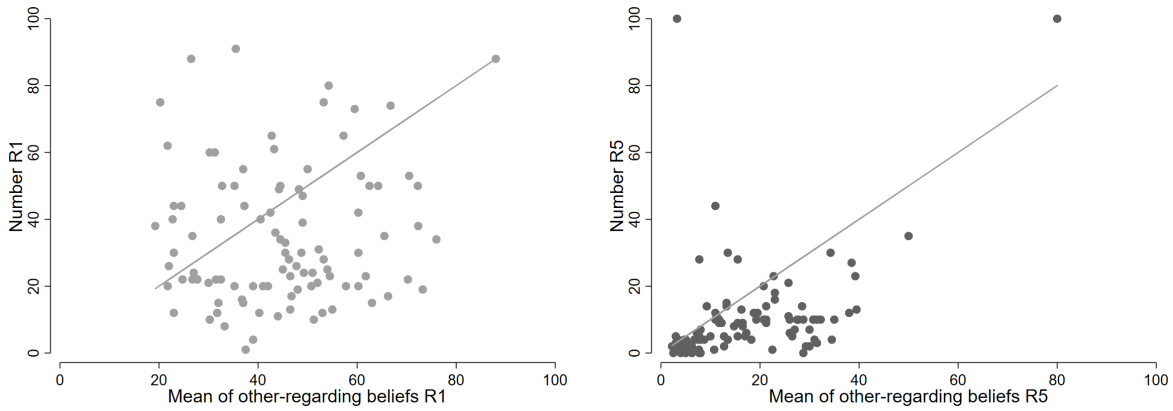

Figure 4.4: Chosen Numbers and Average of Other-Regarding Beliefs

On the other hand, both the actual numbers and mean beliefs decrease and cluster closer to zero in round 5 compared to their outspread distributions in round 1. In fact, the average of stated beliefs before round 1 are significantly higher (signtest, $p \approx 0.00$) compared to round 5.[16] The convergence of observations that can be seen by comparing both rounds represents the approach to the Nash equilibrium.

[16]There are less than 114 observations in Figure 4.4 due to the number of missing observations: 17 and 10 in round 1 and round 5, respectively.

4.2 Depth of Reasoning and Learning

We have shown in Sections 4.1.1 and 4.1.2 that our preliminary results on children's behaviour are in line with the literature on beauty contest games. Most children do not choose the Nash equilibrium in the initial rounds of the game. Hence, zero is not necessarily the winning number. That is, children do not exhibit the rational behaviour that game theory predicts —just like the majority of adults do in previous experimental studies. Instead, the behaviour has been better described by the assumption of bounded rationality. One way of explaining such behaviour is the model of depth of reasoning (or the level-k model, used interchangeably hereafter) first proposed by Nagel (1995) in the context of beauty contest games and widely studied since then (Duffy and Nagel, 1997; Bosch-Domenech et al., 2002; Kocher and Sutter, 2004; Grosskopf and Nagel, 2007; Agranov et al., 2013). Under the assumption of limited rationality individuals employ iterative elimination of dominated strategies in several levels, also known as depths of reasoning or degrees of sophistication. Unlike the assumption of full rationality where only the players who employ an infinite level of reasoning chooses the Nash equilibrium zero, literature on beauty contest games argue that, after some repetitions of the game, subjects are able to reach the Nash equilibrium by employing few levels of reasoning (Sbriglia, 2008).

Our motivation to apply the model of depth of reasoning to the children dataset is primarily to provide further evidence on whether children perform analogously to adults, now within the framework of strategic behaviour. Therefore and also to keep our analysis simple, we follow closely the methodology and notation of Duffy and Nagel (1997) where they examine the behaviour of adults in a median beauty contest game with $p = 1/2$. Since our results, so far, suggested that children at the average age of ten years are capable of playing a beauty contest game, we expect our findings in this section to confirm these results. Hence, we believe that we can achieve meaningful outcomes on children's strategic reasoning processes revealed by their actions in the Goblin Game to the extent that the level-k model explains.

Then, we take the analysis one step forward in an attempt to explore the discrepancy between beliefs and actions we uncovered in Section 4.1.3. Most numbers in the game

were relatively lower than the average beliefs. Since we have information about children's beliefs about their opponents, we pose the question; what levels of reasoning do they think the other players will employ? In other words, we aim to investigate elicited levels of reasoning which we interpret as the other-regarding beliefs on degrees of sophistication.

Finally, based on the results obtained from the model of depth of reasoning, we reconsider the realised numbers within the context of a simple learning model. The objective of studying the learning path of children in the Goblin Game is, first of all, to present another approach to explain their observed behaviour. Moreover, we aim to investigate how the actions of children in a beauty contest game compare to those of adults in terms of learning through experience in a beauty contest game.

4.2.1 The Level-k Model and Realised Levels of Reasoning

In this section, following the literature on beauty contest games and the model of depth of reasoning, we set aside the elicited beliefs of players for the time being and examine only the actions of players, i.e. realised numbers. We begin with a brief description of the model using the characteristics of the Goblin Game: ten rounds of repetition, $p = 1/2$ and common knowledge of the median number after each round. In essence, the model of depth of reasoning investigates whether the players choose the best response according to their underlying beliefs about the behaviour of other players and a reference point.

In the beginning of the game, since there is no information about the behaviour of other players, a reasonable reference point for an inexperienced player would be 50. That is, players choose a random number uniformly distributed within the interval [0, 100], or the most salient number given the game parameters (Nagel, 1995). In the first round, player i is strategic of degree d which solves $x_{i1} = 50p^d$, where x_{i1} is the number chosen by player i. A player is strategic of degree 0 if she exhibits random behaviour, in other words, chooses 50. A player who anticipates all other players are strategic of degree 0 chooses a best response to it by playing $x_{i1} = 50p^1$, 25, and thus, going one step deeper in the process of reasoning to $d = 1$. A player who thinks that all other players are strategic of degree 1, will then best respond by choosing $x_{i1} = 50p^2$, 12.5, and so on. In fact, most children in the first round of the Goblin Game repeatedly chose numbers around 50, 20, 12 (see Figure 4.1) which approximately corresponds to the described levels of sophistication.

In rounds 2-10, the chosen number of each player, the winner and the median number were revealed at the end of the rounds. Hence, players possess information about the actual behaviour of the other players which they can use to adjust their underlying beliefs for the next round. Thus, the median number in the previous round, $median_{t-1}$, serves as the reference point for each player. From the second round and onwards, player i is strategic of degree d if she chooses the number $x_{it} = (median_{t-1})p^d$. Studies that analyse the beauty contest game agree that the behaviour of most players can be classified within the degrees $d = 0, 1, 2, 3$.

To classify the choices of players into discrete levels of reasoning we use intervals around the degrees $d = 0, 1, 2, 3$ with boundaries; $[p^{d+1/2}50, p^{d-1/2}50]$ for the first round and $[p^{d+1/2}(median)_{t-1}, p^{d-1/2}(median)_{t-1}]$ for rounds 2-10. In the first round, the upper boundary for $d = 0$ is set to 50, and choices above 50 are classified into d<0. In the remaining rounds, the upper boundary for d=0 is set to previous round's median number and thus higher choices are classified into $d < 0$. Lastly, chosen numbers that are below the lower bound of $d = 3$ are classified into the category $d > 3$ (Duffy and Nagel, 1997).

Table 4.3 shows the relative frequencies of choices within the defined categories: $d < 0$, $d = 0$, $d = 1$, $d = 2$. $d = 3$ and $d > 3$, with emphasis on the first and second modal frequencies observed in each round. In each of the ten rounds, around 20% of children chose to play dominated strategies ($d < 0$). In the first round, we observe that most of the choices are $d = 1$ and below. In rounds 2-7, again, more than 50% of choices are either $d = 0$ or $d = 1$ and higher levels were not employed as frequently. That is, observing the median number of the previous round, most children choose numbers either equal or a little below the preceding median number ($d = 0$), or they either lower their number to around p times the previous median number ($d = 1$). In rounds 7-8, the increment in the frequencies of $d > 3$ suggests that children employ higher levels of reasoning as the game repeats.

In order to test whether children employ increasing depths of reasoning between consecutive rounds, we create a discrete variable $d\_realised_{it}$ which takes the values of d each player belongs for all ten rounds. We assign the value -1 for the category $d < 0$ and the value 4 for $d > 3$ since they represent at least one degree lower or higher levels of reasoning than the previous degree, respectively. There is no significant difference in degrees of reasoning

between the consecutive rounds 1-2, 2-3, 3-4, 4-5 and the consecutive rounds 6-7, 7-8, 8-9 (signtest, $p > 0.48$). However, we find that levels of reasoning in round 6 are significantly higher than those in round 5 and levels of reasoning in round 10 are significantly higher than those in round 9 (signtest, $p \approx 0.00$).

Table 4.3: Relative Frequencies of Levels of Reasoning Over All Rounds

	Round 1	Round 2	Round 3	Round 4	Round 5
d<0	0.18	0.17	0.18	**0.25**	**0.29**
d=0	**0.21**	**0.29**	**0.25**	0.22	0.25
d=1	**0.39**	**0.41**	**0.42**	**0.38**	**0.29**
d=2	0.16	0.12	0.11	0.11	0.08
d=3	0.02	0.01	0.03	0.02	0.03
d>3	0.04	0.00	0.01	0.03	0.07
	Round 6	Round 7	Round 8	Round 9	Round 10
d<0	**0.21**	0.21	0.22	0.23	0.18
d=0	0.17	**0.27**	**0.29**	**0.30**	**0.19**
d=1	**0.39**	**0.26**	0.19	0.14	0.12
d=2	0.13	0.06	0.06	0.03	0.09
d=3	0.04	0.02	0.00	0.00	0.01
d>3	0.05	0.18	**0.24**	**0.31**	**0.40**

Over the ten rounds of the Goblin Game, the general trend in the levels of reasoning resembles the findings in the literature. We did not find evidence for increasing levels of reasoning between any consecutive rounds from 1-5. However, we see a significant increase from round 5 to 6 and from round 9 to 10, suggesting that children employ higher levels of reasoning after few repetitions of the game. The modal frequencies from round 8 and onwards indicate the same. The major difference of our findings compared to Nagel (1995) and Duffy and Nagel (1997) is that they found the modal frequencies to be within the categories $d = 1$ and $d = 2$ for adult participants. We address this in more detail in Section 5.2.

4.2.2 Elicited Levels of Reasoning

The analysis in this section follows the idea in Lahav (2015) where he found a mismatch between elicited beliefs and actions in beauty contest game experiments. In the Goblin Game, before the first and the fifth rounds, each of the five players stated beliefs about the other players choices. We represent the average belief of each player i about the four

other players with the arithmetic mean. The intention of using the arithmetic mean to summarise other-regarding beliefs is to include the effect of every single belief, without loss of information. By categorising the mean beliefs into degrees of reasoning following the level-k model, we aim to investigate players' beliefs about the levels of sophistication of other players. Since children stated beliefs once before acquiring any information about the other players, the reference point for the average of stated beliefs in round 1 is 50. Accordingly, the reference point for mean beliefs in round 5 is the median number of the previous round, i.e. round 4.

The third column of Table 4.4 shows the relative frequencies of six levels of reasoning measured with elicited beliefs in rounds 1 and 5.[17] In the second column, we repeat the relative frequencies of realised levels of reasoning (measured with chosen numbers) for the purpose of comparison. It can be observed that children, on average, did not state beliefs about other players that equals a strategic degree of 2 or higher. Instead, all their stated beliefs are categorised into $d = 1$, $d = 0$, or $d < 0$. In other words, players think that other players will be strategic of degree 1 or lower in the upcoming round. Looking at the degrees of reasoning corresponding to the numbers played in round 1 and round 5 of the Goblin Game (realised levels of reasoning), we observe that players are categorised into one or more steps higher than their stated beliefs.

To compare the realised and elicited degrees of reasoning we create discrete variables $d\_elicited_{i1}$ for mean beliefs round 1 and $d\_elicited_{i5}$ for round 5 using the same method described in Section 4.2.1. The variables take the value of the degree of reasoning each child exhibits, while representing $d < 0$ with -1 and $d > 3$ with 4. Realised levels of sophistication are significantly higher than the elicited levels of sophistication, both in round 1 and in round 5 (signtest, $p \approx 0.00$).

Figure 4.5 shows the differences between the d variables that represent the realised and elicited levels of sophistication, $(d\_realised_{it} - d\_elicited_{it})$ for $t = 1$ and $t = 5$. On both graphs, the differences are skewed to the right with 70% and 73% of the differences being equal or higher than one in round 1 and round 5, respectively. In round 1, 22% of the children were one step ahead in terms of their actions compared to their beliefs about

[17] The missing observations in the belief elicitation questions were taken into account when calculating the categorisation and relative frequencies in order to avoid missing observations to be treated as any of the categories. The sample sizes are 97 and 104 in round 1 and round 5, respectively.

Table 4.4: Relative Frequencies of Elicited Levels of Reasoning

Round 1	Realised	Mean Belief
d<0	0.18	**0.34**
d=0	**0.21**	**0.36**
d=1	**0.39**	0.30
d=2	0.16	0.00
d=3	0.02	0.00
d>3	0.04	0.00
Round 5	Realised	Mean Belief
d<0	**0.29**	**0.78**
d=0	**0.25**	**0.19**
d=1	**0.29**	0.03
d=2	0.08	0.00
d=3	0.03	0.00
d>3	0.07	0.00

others. In the fifth round, 27% of them employed one step higher level of sophistication than their stated beliefs.

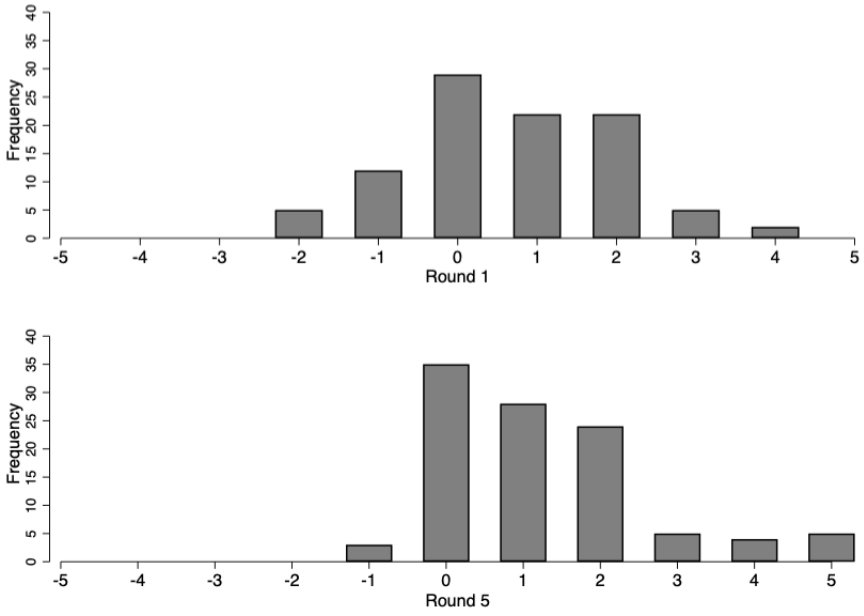

Figure 4.5: Differences Between Realised and Elicited Depths of Reasoning

The findings of this section confirm the relationship we observed in Section 4.1.3 between elicited beliefs and actions. Realised levels of sophistication are higher than the elicited levels of sophistication, which translates to choices being lower than the stated beliefs about others. The fact that players adjust their behaviour downwards when actually

playing the game can be interpreted as the behaviour being in line with the theory behind the model of depth of reasoning. The theory implies that individuals respond to their underlying beliefs about others. Camerer et al. (2004) argue that the goal in the beauty contest game is to be one step ahead of opponents. Nevertheless, even if players consciously act one step ahead of their beliefs about others, this may not lead them to win the game when the beliefs they hold about their opponents are not accurate.

4.2.3 A Simple Learning Model

Sections 4.1.2 and 4.2.1 documented that chosen numbers in the Goblin Game gradually converge to the Nash equilibrium as the game repeats. However, the levels of reasoning that players employed do not increase significantly in any consecutive rounds until the fifth round. Thus, in this section, we aim to examine whether a simple learning model can better explain the behaviour of children in the Goblin Game. To that end, we apply a qualitative model of adaptive learning behaviour, known as the learning direction theory, first proposed by Selten and Stoecker (1986) and adapted to the beauty contest game by Nagel (1995) and Duffy and Nagel (1997). The empirical literature on the application of learning models to experimental beauty contest games has evolved since the publication of the representative papers mentioned above. Regardless, we choose to study the learning direction theory, since our primary purpose is to highlight similarities and contrasts of our results compared to the literature on experimental beauty contest games. The intuitive and straightforward nature of the learning model serves to keep our analysis simple and it also allows us to provide a secondary approach which may reveal further insights on the strategic behaviour of children.

The learning direction theory rests on the question of whether the behaviour of interest adjusts towards the desired direction over time due to past experience. In the context of beauty contest games, the theory predicts that a player who observes her choice in the current round, t, to be higher (lower) than the winning number of the round, will decrease (increase) her adjustment ratio in the next round, $t+1$. The adjustment ratio of round t is defined by the ratio of the chosen number in t, $x_{i(t)}$, to previous round's median, $\frac{x_{i(t)}}{median_{(t-1)}}$. Accordingly, the adjustment ratio of round $t+1$ is defined by $\frac{x_{i(t+1)}}{median_{(t)}}$. The behaviour that is consistent with the predictions of learning direction theory from

round t to round $t+1$ is summarised below and otherwise considered inconsistent. That is, a player who observes her chosen number is above p times the median in round t, will decrease her adjustment ratio in the next round and vice versa,

$$x_{i(t)} > p * median_{(t)} \Rightarrow \frac{x_{i(t+1)}}{median_{(t)}} < \frac{x_{i(t)}}{median_{(t-1)}}$$

$$x_{i(t)} < p * median_{(t)} \Rightarrow \frac{x_{i(t+1)}}{median_{(t)}} > \frac{x_{i(t)}}{median_{(t-1)}}$$

In our analysis, we allow for a third way of consistent behaviour. That is, keeping the adjustment ratio constant from round t to $t+1$, if the chosen number in round t equals the winning number,

$$x_i(t) = p * median(t) \Rightarrow \frac{x_i(t+1)}{median(t)} = \frac{x_i(t)}{median(t-1)}$$

Finally, we exclude the sessions where the median number was zero for the following reasons. Firstly, under the model of learning direction theory, it is infeasible to calculate the adjustment ratio when the median number equals zero. Moreover, if the median number equals zero, the winning number also equals zero. Thus, it is not possible for any player to have selected a number below the winning number, making it challenging to examine the theoretic direction of their adjustment behaviour. Secondly, the median number being zero means that at least three or more players in the session have selected zero. Observing that it occurs more often in the later rounds of the game, we interpret this behaviour as players have reached the Nash equilibrium. Hence, there is no room for considering their learning path.[18]

Table 4.5 demonstrates the relative frequencies of behaviour over the ten rounds classified into to the possible paths of learning direction theory. The behaviour consistent with the theory from round t to $t+1$ is underlined and the cumulative consistent frequencies appear in bold font. That is, players who update their adjustment ratio downwards (upwards) from one round to the next when the chosen number was above (below) p times

[18]The sample sizes used when calculating the relative frequencies in Table 4.5; Full sample (N=114) for rounds 1-2, 2-3, 3-4 and 4-5. 109, 104, 104, 94 and 79 for rounds 5-6, 6-7, 7-8, 8-9 and 9-10, respectively. Nagel (1995) excludes the winners regardless of the winning number being equal to zero or a positive number. We exclude the sessions only when the winning number equals zero and allow for a third category in the classification.

the median number of the current round. Also, players that keep their adjustment ratio constant when the chosen number was equal to p times the median number. In most transitions from round t to $t+1$, more than 50% of all players show behaviour that is in line with the predictions of learning direction theory (except rounds 4-5, 6-7 and 7-8). Until round 5 (the first three transitions) 55% of children, on average, exhibited consistent behaviour, while over all nine transitions, the corresponding percentage was 54% of the children.

Duffy and Nagel (1997) finds the behaviour of 72% and 62% of the players in their $1/2 - median$ game consistent with learning direction theory, over ten rounds and over the first four rounds respectively.[19] The frequencies of behaviour in the Goblin Game that are consistent with the theory are lower than those in the literature. Even so, we interpret more than half of the children acting consistently as suggestive evidence in favour of the predicted direction of the learning model. That is, more than half of the players throughout the game adjusted their choices towards the desired direction depending on their individual experience in the previous round.

4.3 Determinants of Game Performance

Previously, we demonstrated that children are able to perform in a meaningful way in the Goblin Game. Thus, we take the analysis further to investigate possible determinants of the game performance. The analysis evaluates the relevance of certain elements to the game performance of children within a regression framework.

4.3.1 Measures of Performance

The rules of the Goblin Game specify that the player closest to half the median of all chosen numbers wins the round and receives a coin. In case of a tie, all winners receive a coin and, children are encouraged to win as many coins as possible over the ten rounds. Hence, children with greater amounts of coins at the end are the ones that showed better performance in the game. We consider the total number of coins as a direct measure of game performance. However, the cumulative number of coins has limitations. It is a

[19] The corresponding percentage in Nagel (1995) is on average 73% over four repetitions.

Table 4.5: Relative Frequencies of Behaviour Classified According to the Learning Direction Theory

		Rounds 1-2	**Rounds 2-3**	**Rounds 3-4**
$x > p * median$	Decreased	0.44	0.48	0.49
	Increased	0.41	0.40	0.40
	Constant	0.00	0.02	0.02
$x < p * median$	Increased	0.11	0.09	0.04
	Decreased	0.01	0.01	0.02
	Constant	0.00	0.00	0.00
$x = p * median$	Constant	0.00	0.00	0.00
	Increased	0.02	0.00	0.02
	Decreased	0.01	0.00	0.01
	Consistent	**0.55**	**0.57**	**0.53**
		Rounds 4-5	**Rounds 5-6**	**Rounds 6-7**
$x > p * median$	Decreased	0.40	0.57	0.40
	Increased	0.45	0.22	0.42
	Constant	0.02	0.02	0.03
$x < p * median$	Increased	0.08	0.09	0.06
	Decreased	0.02	0.02	0.05
	Constant	0.00	0.02	0.01
$x = p * median$	Constant	0.00	0.00	0.00
	Increased	0.03	0.05	0.03
	Decreased	0.00	0.01	0.00
	Consistent	**0.48**	**0.66**	**0.46**
		Rounds 7-8	**Rounds 8-9**	**Rounds 9-10**
$x > p * median$	Decreased	0.36	0.46	0.48
	Increased	0.35	0.23	0.19
	Constant	0.10	0.12	0.14
$x < p * median$	Increased	0.11	0.10	0.06
	Decreased	0.02	0.00	0.04
	Constant	0.04	0.03	0.05
$x = p * median$	Constant	0.01	0.01	0.00
	Increased	0.01	0.04	0.01
	Decreased	0.00	0.01	0.03
	Consistent	**0.48**	**0.57**	**0.54**

discrete measure that only defines the winner of the game, and does not account for being close to the winning number. Players who fail to earn a coin with a short distance to the winning number are treated as equal to the ones with larger distances. In order to distinguish such players, we employ a continuous measure of performance in addition to the total number of coins: *distance to the best response.*

Evaluating performance (posteriori, after playing) by the winning rule of the Goblin Game (the closest player to p times the median wins the game), would bring us to the conclusion that children with shorter distances to half the median have superior performance. Thus, the best strategy would be to select it. However, choosing p times the median (apriori, before playing) would only make the players that are posteriori below the median the best performers.[20] For those subjects with numbers equal to or greater than the median number, selecting p times the median instead of their number would move the winning benchmark downwards. In fact, it would position them in the place of the median. Therefore, for the players who have numbers equal or above the median, the best response would no longer be equal to half the median but to half of the second lowest number among the choices of all players. On the contrary, for individuals with numbers smaller than the median, choosing p times the median would not alter the median and hence the winning rule would remain as the best response. In other words, we define the best response z_i as:

$$z_i = \begin{cases} \text{second lowest number}/2, & \text{if number}_i \geq \text{median} \\ \text{median}/2, & \text{otherwise} \end{cases}$$

Now that we have identified the best response for the game, we estimate its absolute distance to the choices over rounds. The distance to the best response, z_i, evaluates more precisely how children anticipate, apriori, the behaviour of others and how they react to it.

Intuitively, the total number of coins and the distance to the best response are oppositely related measures of performance. Players with shorter distances to the best response have higher chances of earning a coin. Figure 4.6 shows how the average absolute distances to the best response develops over time for three levels of performance measured in coins. As can be seen, those who acquired most coins at the end of the game (7-9 coins) have lower

[20] Assuming that the choices of the other players remain constant.

distances compared to the ones who acquired the least coins (0-3 coins). On the other hand, regardless of the number of coins, average distances decrease over time indicating that players choose numbers closer to the best response as the game repeats. Children who won less than seven coins in total, have a steeper decrease in their distances to the best response (approximately until rounds 5-6). That means experience from the initial rounds has a favourable and relatively greater impact on this group compared to the ones that won seven coins or more.

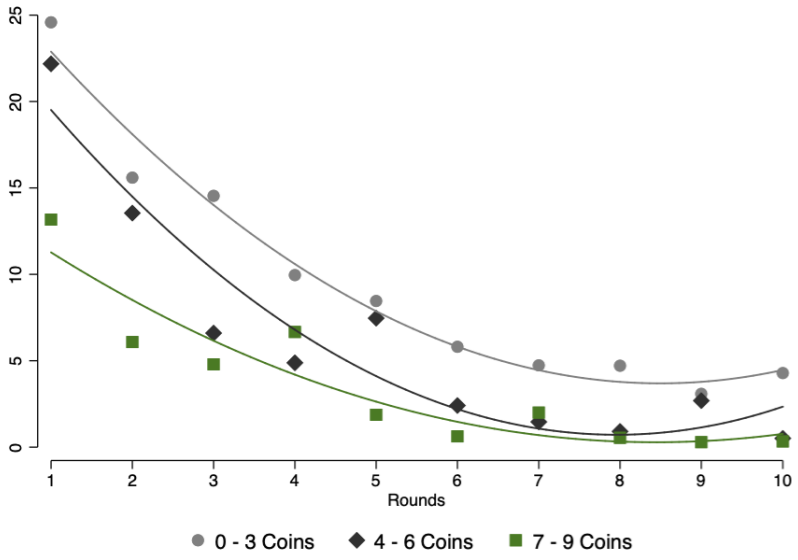

Figure 4.6: Average Absolute Distances to The Best Response Over Rounds

In order to summarise the distances to the best response we first square the continuous variable. The reason for this is that equal distances to best response (in absolute terms) cannot be treated as equivalent since the measure of performance must distinguish whether the best response is close or far away from the Nash equilibrium. Afterwards, in order to make the measure comparable over rounds, we standardise it and add up from round 1 to round 10 to create one observation per child that represents the overall performance in terms of proximity to the best response.

Distributions of both measures of performance in Figure 4.7 show that children earn on average four coins throughout the Goblin Game. Over all rounds, the mean distance between choices and best responses is 2.5 standard deviations. The distribution of this variable appears skewed to the left. The right tail of the distribution consists of children with more than three standard deviations away from the best response, corresponding to

approximately 20% of the observations.

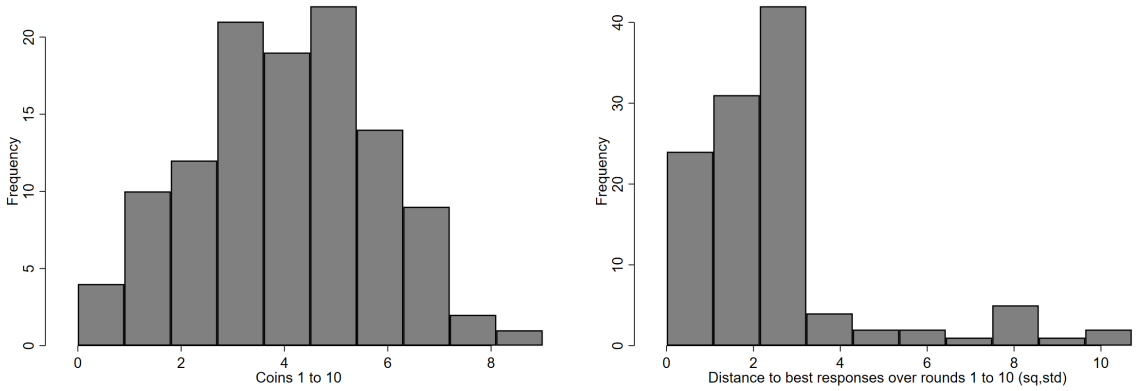

Figure 4.7: Measures of Game Performance

The correlation between the two measures of performance, -0.17, is weak but significant at a 10% level. The negative relationship reflects the fact that children with larger amount of coins at the end of the game were the ones who deviated less from the best response. Nevertheless, in the absence of a strong relationship between the measures, one cannot assume that, the players who earned more coins were also closer to choose the best response throughout the game. In the regression analysis we include both measures of performance as dependent variables, since they provide different insights. A caveat on distance to the best response is that, while capturing more details about the behaviour of the participants the variable may be noisy and lack predictive power due to the limited number of observations.

Before moving to the analysis of potential predictors of the game performance, we would like to mention several points about the methodology. We employ OLS regressions with standard errors clustered at the session level to control for common unobserved elements among participants within the same session. Although children were randomly assigned into sessions to play the Goblin Game, the data may not be independent across observations. Since sessions were formed with children from the same classroom, the dynamics within every session may also affect the game performance. Further reasons can include the way children were seated, friends and environment in general. Hence, the residuals in the same session may not be independent of one another. In addition, we control for session fixed-effects to take into account the variations across groups.

In terms of control variables we include the following child-level characteristics: age,

gender, math skills and whether children attend to the fifth grade. Besides our variables of interest, these characteristics are expected to affect the game performance. It is of special interest to include age as a control because participants of the Goblin Game were within the age range of 8-11. The literature presents evidence on the impact of age from early childhood to adolescence in strategic situations (Brosig-Koch et al., 2015; Brocas and Carrillo, 2018a,b). Therefore, we account for common characteristics of each age that could affect the performance of children in the game. Further, the literature on strategic reasoning and rational behaviour report mixed results on gender (Czermak et al., 2016; Brocas and Carrillo, 2018b). As an additional control we include math skills proxied by the evaluations of teachers. Higher grades in mathematics can be significantly relevant in strategic settings (Czermak et al., 2016).

4.3.2 Understanding of the Game

Comprehension of the game rules is of the utmost importance in experimental studies with children participants. In order to avert the concerns in this regard, researchers adopted distinctive approaches in experimental designs. Sutter et al. (2019) mention the widely applied procedural features concerning the understanding of the game that implemented extensively to date, such as explaining the rules to children in a one-to-one setting, including control questions to make sure the rules are accurately understood and conveying the instructions in a standardised fashion. The Goblin Game study with children fulfils the procedures far strictly on that note. As detailed in Section 3.1, the precise design of explaining the game rules is a vital part of the study. The rules were conveyed to children repeatedly in several different ways (story of the goblin, game rules in five-steps). Moreover, the game itself was designed to ensure comprehension at the highest possible extent with physical game material (game board, pawns, coins) and simplified game parameters: the third player (the median) and halfway back (p=1/2).

Although precise levels of understanding are unobservable, the study design includes questions about comprehension rated by interviewers and children regarding other players, both before and during the game. We expect the assessment of understanding by others to proxy the unobservable understanding levels. This allows us to investigate the understanding of the game and how it relates to the game performance. Interviewers,

first of all, asked children to explain back all the five steps of the game rules and then evaluated their explanations from 1 (not completely understood) to 4 (instantly, completely and correctly explained). The scores are positively biased, 79% of the subjects got the highest ratings which provides little variation. Thus, the stepwise rule assessment has limitations to anticipate the game performance as we observe in the first row of Table 4.6.

Table 4.6: OLS Estimates of Ratings of Understanding

	(1) Coins R1-R10	(2) Distance to best response R1-R10	(3) Coins R5-R10	(4) Distance to best response R5-R10
Interviewer ratings on step-wise game rules	0.134 (0.171)	-0.074 (0.299)	-0.025 (0.144)	0.093 (0.218)
Interviewer ratings on understanding R1	0.073 (0.134)	-0.668*** (0.212)		
Children ratings on understanding R1	-0.369 (0.363)	0.292 (0.397)		
Interviewer ratings on understanding R5			1.016*** (0.160)	-0.574** (0.261)
Children ratings on understanding R5			-0.165 (0.131)	-0.123 (0.289)
Control variables	✓	✓	✓	✓
Session FEs	✓	✓	✓	✓
Observations	111	111	111	111
R^2	0.441	0.245	0.714	0.328

Notes: All the independent variables are standardised ratings of interviewers and children. The vector of control variables includes a dummy that takes the value of 1 when child i is female and 0 otherwise; standardised teacher ratings on a 4-point scale (1 = poor, 4 = excellent) about the math skills of child i; a dummy variable which takes the value of 1 if a subject attends the 5th grade; and the age of participants in years, squared to account for non-linear relationships with the game performance. The control variable of teacher-rated math skills has missing values in three observations, thus the sample size reduces to 111. Standard errors clustered at the session level are in parentheses. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Moving forward, Table 4.6 presents average ratings of four interviewers based on their general impression and ratings of children on other players about the understanding of the game. Both evaluations were made once before the game (round 1) and once again during the game (round 5). Our objective is to study the relationship between these evaluations and the game performance in future rounds. To avoid the influence of past rounds in our results, we only include in the dependent variables the remaining rounds from the time that ratings were done.

The interviewer ratings on understanding before the game lack relevance in terms of winning frequency but it is significantly related to best responding behaviour. In round

5, interviewers ratings also predict the proximity to the best response of children and furthermore became strongly related to the number of coins earned from round 5 to round 10. The ratings of interviewers in round 5 indicate that higher understanding scores yields to more coins and shorter distances to the best response. One standard deviation increase in interviewer ratings is related to an increment of 1.02 standard deviations in the number of coins earned from round 5 to 10. Likewise, interviewer ratings also associates with 0.57 standard deviations decrease in the distance to the best response. On the other hand, variables of understanding measured from children's perspective lack predictive power. The estimates for these variables anticipate neither the accumulated number of coins nor the best response behaviour in any of the rounds.

One may argue that the evaluations in round 5 are reflecting the past winners instead of accurately estimating the future performance. The insignificant and weak correlations between the performance measures calculated for different parts of the game[21] suggest that this is not the case. If interviewer ratings reflected the past winners, we would not expect them to be significant in performance measures for rounds 5 to 10. However, reflecting the past winner could be the case when children rated the understanding of other players. Indeed, as we observed previously, children tend to base their perceptions on past winners. Therefore, their evaluations of understanding can be influenced by the performance in previous rounds.

To summarise, the findings suggest that average ratings of four interviewers relate to the best responding behaviour of children. After some rounds of experience, interviewers were also able to identify understanding features that are related to the winning frequency. In contrast, the perception of children about other players' understanding of the game does not relate with the measures of performance. Thus, our results point out that the ratings of external observers are more reliable to describe the behaviour of players in the game. Although, it is important to keep in mind that players in the Goblin Game were children and external observers were trained researchers.

[21]The pairwise correlations between coins R1-R4 and coins R5-R10 is 0.11 and insignificant. The distance to the best response in R1-R4 is also weakly and insignificantly correlated to the distance to the best response in R5-R10 (0.10).

4.3.3 Cognitive Ability and Empathy

On one hand, the beauty contest game requires mental effort to estimate p times the median or the mean of the chosen numbers. In our case, the Goblin Game design simplifies this process in order to adapt the game to children. On the other hand, successful performance in the beauty contest game not only requires strategic reasoning but also the ability to take the perspective of others. According to the theory individuals anticipate the movements of others to form beliefs about their behaviour and use them as a reference for their own decisions. Brocas and Carrillo (2018b) argue that children develop the ability to be aware of the differences that exist between one's point of view and that of others between the ages 4 to 7. The ages of the subjects in our sample are above this threshold.

In this regard, we attempt to investigate the relevance of cognitive ability and empathy to the game performance. As mentioned in previous sections, children's questionnaires consisted of the Raven test to measure fluid IQ, FEAS scale and IRI test as self-rated measures of empathy. In addition, teachers rated the empathy skills of their students and the interviewers conducted two behavioural tasks with children to proxy their empathy skills: E on the forehead and contagious yawning.[22]

Table 4.7 presents how fluid IQ and empathy measures are distributed. All variables except the dummy for the empathy task are standardised to mean zero. Other than the FEAS scale, the distributions are quite similar to each other with values of up to three standard distributions below the mean and with about one above the mean. The FEAS scale is skewed to the left compared to the other measures of empathy. Moreover, 22% of children drew an "E" readable from the interviewer's perspective.

Table 4.8 displays the results of the OLS regression with various measures of empathy and fluid IQ on the game performance. We observe that fluid IQ represented by the Raven test is not significantly correlated to both measures of game performance. This finding stands in contrast to the evidence in Gill and Prowse (2016) where authors found that the cognitive ability measured by the Raven test is relevant to the behaviour in the beauty contest game.

[22]We do not include the yawning task in our analysis because only 11 children out of 114 displayed contagious yawning. The variable lacks representation within the sample.

Table 4.7: Descriptive Statistics: Cognitive Ability and Empathy Skills

	Mean	SD	Min	Max	N
Fluid IQ (Raven)	0	1	-3.4	1.2	114
Empathy (FEAS)	0	1	-6.0	1.2	114
Empathy (IRI)	0	1	-3.1	2.4	114
Teachers' empathy rating	0	1	-2.5	1.2	112
Empathy Task	0.22	0.42	0	1	112

Note: The table reports standardised values of the fluid IQ (Raven), Empathy (FEAS) and Empathy (IRI) test scores. The table also includes standardised values of teachers' evaluations of empathy. Empathy task stands for the "E on the Forehead" task.

Table 4.8: OLS Estimates of Cognitive Ability and Empathy Skills

	(1) Coins R1-R10	(2) Distance to best response R1-R10
Fluid IQ	0.114	-0.625
	(0.360)	(0.374)
Empathy (IRI)	-0.612***	0.359
	(0.145)	(0.263)
Empathy (FEAS)	0.232*	-0.277
	(0.124)	(0.259)
Teachers' empathy rating	0.368	-0.0788
	(0.228)	(0.201)
Empathy Task	1.366***	-1.044
	(0.485)	(0.793)
Control variables	✓	✓
Session FEs	✓	✓
Observations	109	109
R^2	0.548	0.292

Notes: All the independent variables are standardised scores of the tests and the empathy task is a dummy that takes the value of 1 if the child drew an "E" on her forehead readable from the interviewer's perspective and 0 otherwise. The vector of control variables includes a dummy that takes the value of 1 when child i is female and 0 otherwise; standardised teacher ratings on a 4-point scale (1 = poor, 4 = excellent) about the math skills of child i; a dummy variable which takes the value of 1 if a subject attends the 5th grade; and the age of participants in years, squared to account for non-linear relationships with the game performance. The control variable of teacher-rated math skills has missing values in three observations and the teachers' empathy rating in two, thus the sample size reduces to 109. Standard errors clustered at the session level are in parentheses. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

While the FEAS scale and the empathy task are significantly and positively related to the winning frequency, teacher-rated empathy is not. This might be explained by the lack of variation in teachers' ratings; 74% of children received maximum scores of empathy. The results for the empathy task suggest that those who drew an "E" readable from the interviewer's perspective, on average, won 1.3 more coins by the end of the game. A higher score of IRI displays a strong correlation with the total number of coins. However, the IRI test scores show an opposite relationship with both measures of performance contrary to every other measure of empathy. Successful performance in the Goblin Game is associated with a higher number of coins and shorter distances to the best response which is not the case for higher scores on the IRI test.

Measures of empathy and fluid IQ do not relate to the game performance at a significant level in terms of proximity to the best response. Despite this fact, the trends we observe for each relationship are consistent across both measures of performance. In short, the suggestion we derive from our results in this section is that, empathy provides insights associated with the performance in the Goblin Game. This can be attributed to the nature of the beauty contest games, which requires players to take the perspective of others to perform well.

4.3.4 Stated Beliefs

Best Responding to Stated Beliefs

Most studies on beauty contest games are based on the assumption that choices of the players represent their beliefs. As a result, literature on this topic largely based its conclusions on actions and beliefs remained unobserved. In order to overcome this limitation, elicitation methods are applied to transform the beliefs into observables in the laboratory. As explained in previous sections, children in the Goblin Game were required to state the numbers that represent their beliefs about the others before round 1 and before round 5. Belief elicitation in our case always comes before making a decision in the game. The process might have encouraged children to think more carefully about the choices of their peers in the upcoming round. In Sections 4.1.3 and 4.2 we discussed how choices of children compare to their stated beliefs. In fact, children on average adjust their choices downwards compared to the average beliefs they stated about others.

Nevertheless, it remains uncertain if children used the behaviour of their opponents as a guide when making their decisions. We cannot discard the possibility that children might have considered instead, a different heuristic or a learning rule independent of beliefs to make choices in the game.

The purpose of this section is to explore best responding behaviour to the stated beliefs and how this relates to the performance in the game. We measure best responding to stated beliefs with a dummy variable that equals one if a subject chooses a number that results with the shortest distance to the best response of her stated beliefs. In other words, if the chosen number of the player is the closest to half of the second lowest belief (the number that would made her win) out of the four that she stated regarding other players. Less than half of our subjects best responded to their own stated beliefs in rounds where belief elicitation took place, 37% and 40% respectively. Evidence presented in the review of Schotter and Trevino (2014) on studies with belief elicitation applied in the laboratory indicates higher numbers[23] of players best responding to their stated beliefs compared to children in the Goblin Game. However, we do not consider this as a direct contrast to our finding since results largely depend on the method of belief elicitation and the game itself. In a game where the majority of players did not follow such a strategy, our objective is to investigate how favourable it was (in terms of performance) for the players to act consistent to their stated-beliefs.

Accuracy of Stated Beliefs

If players' stated beliefs do not capture the true behaviour of their opponents, having consistent choices to these beliefs may not be enough to target the best response in the game. Hyndman et al. (2012) argue that individuals with more accurate beliefs about others, best respond more often. We consider that the relationship also depends on the ability of the subjects to make accurate predictions about the behaviour of their opponents. Therefore, we introduce a variable that measures the accuracy of the stated beliefs. Since it would be unrealistic to expect precise guesses from the players in our setting, we measure the accuracy of their stated beliefs in terms of distance. Our measure of belief accuracy is a continuous variable that captures the difference between the stated beliefs of subject i

[23]Percentages of subjects who best responded to their stated beliefs according to previous studies: Nyarko and Schotter (2002) 75%, Danz et al. (2012) 63%, Hyndman et al. (2013) 62% and Manski and Neri (2013) 89%, to mention a few.

about each of the other four players and the corresponding actual choices of her peers.

Since every child made four predictions about the choices of others, we ended up with four distances per child both before round 1 and before round 5. Then, we squared the distances to penalise the observations that are further apart and standardised them to make the measure comparable over rounds. In order to acquire a cumulative measure of belief accuracy we added up the squared and standardised distances of every subject. Intuitively, larger distances between a player's stated beliefs and the chosen numbers of her opponents means that her predictions were less accurate.[24] In round 1, 32% of the calculated distances were within one standard deviation around the mean, while this percentage increased to 51% in round 5. Thus, the measure of accuracy reveals that children improved their predictions as the game repeated. Indeed, Camerer et al. (2002) considered repeated games to study the accuracy of stated beliefs and concluded on the relevance of experience in this regard.

Merely studying the accuracy of stated beliefs is of little value if we do not consider how players respond to those predictions when making decisions in the game. Therefore, we concentrate on the combined effect of those variables on the game performance by adding interaction terms to the regression model. The interactions allow us to observe the effect of making accurate predictions about the behaviour of opponents depending if the subject best responds to those predictions.

Table 4.9 displays the regression estimates for variables of best responding to stated beliefs, inaccuracy and their interactions on the measures of game performance. Having inaccurate beliefs alone seems not to be detrimental to the game performance looking at the weak and ambiguous correlations. In the case children report inaccurate beliefs and best respond to them, the effect on the winning frequency becomes significant. The signs of the interaction terms for both rounds reveal that having inaccurate beliefs and best responding to them leads to fewer coins in the game. The variables together are strong enough to anticipate a decrease of 0.37 and 0.44 standard deviations in the total number of coins (significant at a 10% level). Regarding the second measure of game performance, the estimates of best responding to inaccurate beliefs are not reliable. However, the direction of the relationship is in line with what we observe in terms of coins.

[24] In order to avoid confusion, we name the variable "inaccuracy" of beliefs in the regression analysis.

Table 4.9: OLS Estimates of Best Responding to Stated Beliefs and Inaccuracy

	(1) Coins 1 to 10	(2) Distance to best response 1 to 10
Best responding to stated beliefs R1	-0.077	1.414*
	(0.505)	(0.741)
Best responding to stated beliefs R5	-0.266	0.518
	(0.341)	(0.657)
Inaccuracy R1	0.087	-0.259**
	(0.195)	(0.110)
Inaccuracy R5	-0.062	-0.518
	(0.108)	(0.657)
Best responding to stated beliefs R1 × Inaccuracy R1	-0.373*	0.394
	(0.207)	(0.240)
Best responding to stated beliefs R5 × Inaccuracy R5	-0.447*	0.240
	(0.238)	(0.313)
Control variables	✓	✓
Session FEs	✓	✓
Observations	90	90
R^2	0.486	0.342

Notes: The independent variables are squared and standardised distances except for *Best Responding to Stated Beliefs* which is a dummy variable. The vector of control variables includes a dummy that takes the value of 1 when child i is female and 0 otherwise; standardised teacher ratings on a 4-point scale (1 = poor, 4 = excellent) about the math skills of child i; a dummy variable which takes the value of 1 if a subject attends the 5th grade; and the age of participants in years, squared to account for non-linear relationships with the game performance. The control variable of teacher-rated math skills has missing values in three observations. The rest of the missing observations are produced by players who did not complete the questionnaires regarding the beliefs. Standard errors clustered at the session level are in parentheses. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5 Adults

In this section, we replicate the analysis carried out for children with the sample of adults explained in Section 3.2. The purpose of this complementary study is to provide a foundation for further comparison with a control sample of adult subjects within the same framework. To begin with, we describe the general behaviour of adults in the Goblin Game. Then, proceeding in parallel with our study on children, we apply the models of depth of reasoning and learning direction theory in order to deepen the analysis on actions and stated beliefs. Lastly, we investigate how certain attributes of adult players relate to their performance in the Goblin Game.

5.1 Descriptive Statistics

Behaviour in Round 1

Figure 5.1 shows the distribution of choices in the first round. The mean choice of adults is 24 and the median is 21 with a standard deviation of 15. All of these values are lower compared to children in the first round of the Goblin Game. Even so, only 4% of the adult players chose numbers below 10. The concentration of the tallest bars around numbers between 15 and 25 indicate that most decisions lie within this range, with 17 being the mode choice. Unlike the first round behaviour of children, dominated strategies (numbers larger than 50) are not played as frequently by adults in the first round of the game (6%).

Behaviour in Rounds 2 to 10

Figure 5.2 presents the frequencies of the chosen numbers for all ten rounds of the Goblin Game.[25] Choices are grouped into intervals of five, except for the last category which contains all dominated strategies (50:100]. The rounds with the highest frequency of numbers below five are 5, 6 and 7, with the sixth round having the highest number of observations (103 out of 120 participants). The visualisation of all choices grouped together allows us to observe the convergence towards the Nash equilibrium. Observations of adults

[25] A more detailed visualisation of the chosen numbers for each round of the Goblin Game can be found in Appendix A.

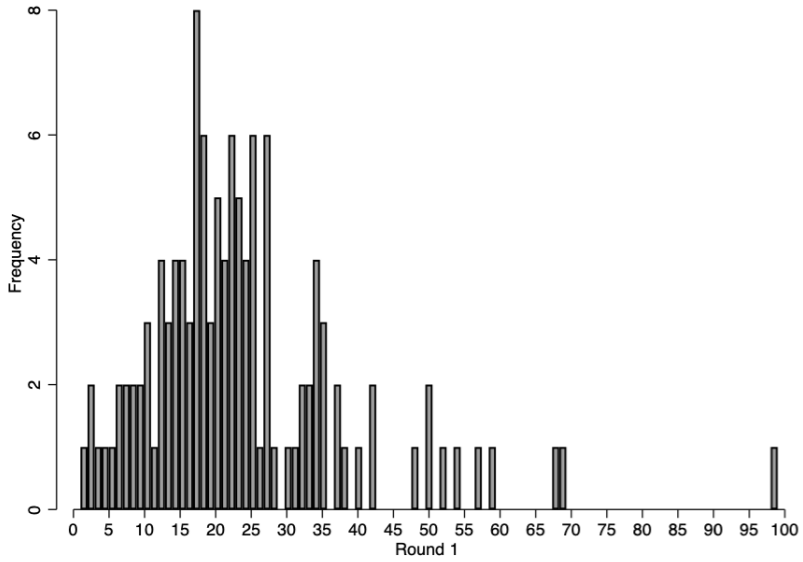

Figure 5.1: Chosen Numbers in Round 1

are more concentrated around the modal choices, while choices of children exhibited a wider dispersion.

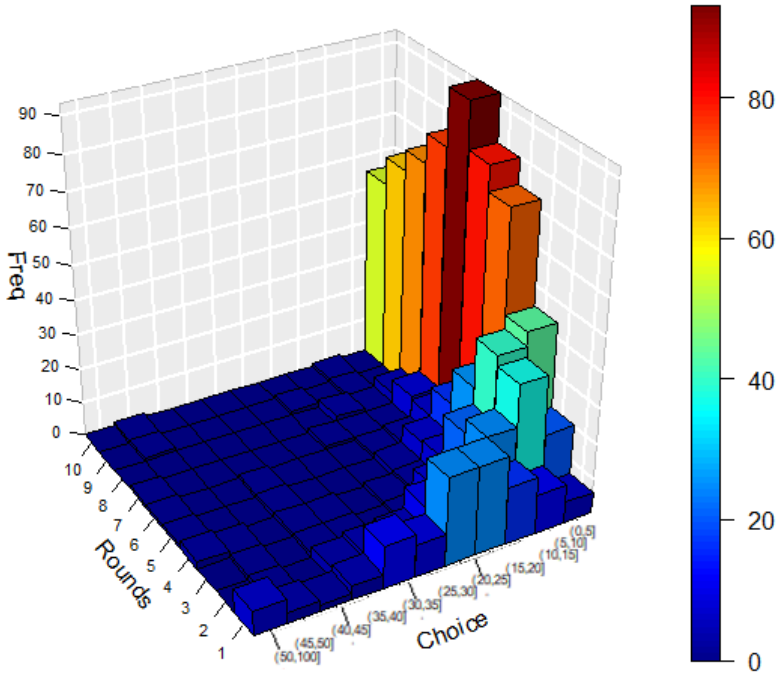

Figure 5.2: Chosen Numbers in the Goblin Game

Figure 5.3 complements the previous graph by displaying the transitions of choices from round t to round $t+1$. The sample of adults repeats the trend observed in children; the chosen numbers constantly decline over time. Indeed, there are significantly more observations below the diagonal in any of the consecutive rounds (signtest, $p \approx 0.00$). In

contrast to the corresponding graph for children, the transitions of adults show a sharper convergence to the Nash equilibrium. For instance, in the tenth round, 90% of choices are between the numbers 0 and 1 whereas only 68% of the children reach these numbers in the final round. A common feature in both graphs is the persistent presence of outliers.

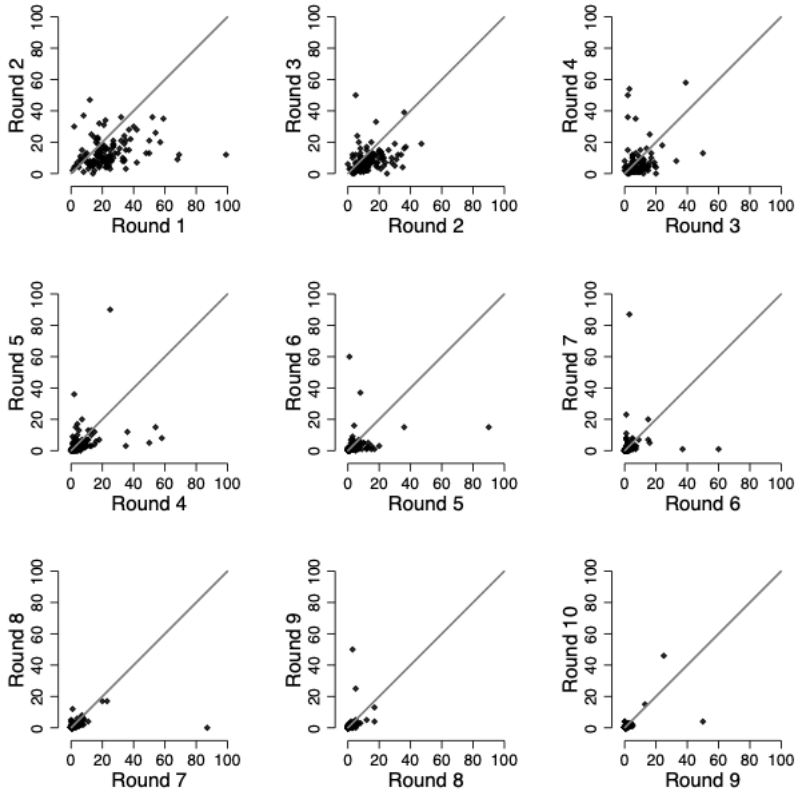

Figure 5.3: Transitions of Chosen Numbers from Round t to Round $t+1$

Table 5.1 reports the medians and means of the choices of players in each of the rounds along with their percentage decrease rates from round $t-1$ to round t. The median of choices depart from a lower number in adults (21) compared to children (28) in the first round of the game. Both indicators decrease between consecutive rounds and the medians for adults are lower than those of children in all ten rounds. Compared to choices of children, adults approach the Nash equilibrium earlier in the game.

Stated Beliefs

The belief elicitation method of adults differs in a few ways compared to the procedure carried out with children. First and most important, adults received a payoff for the accuracy of their beliefs. Secondly, the questions clearly asked the participants not to

Table 5.1: Medians and Means of Chosen Numbers Over Rounds

	Median	1-Median(t)/Median(t-1)	Mean	1-Mean(t)/Mean(t-1)
Number Round 1	21		23.64	
Number Round 2	11	0.48	13.47	0.43
Number Round 3	7	0.36	8.31	0.38
Number Round 4	4	0.43	6.64	0.20
Number Round 5	3	0.25	4.83	0.27
Number Round 6	2	0.33	3.13	0.35
Number Round 7	1	0.50	2.75	0.12
Number Round 8	1	0.00	1.59	0.42
Number Round 9	1	0.00	1.54	0.03
Number Round 10	0	1.00	1.13	0.27

specify a number for themselves. In addition, the questionnaires did not collect information on the belief about the expected winner and the perceived understanding of the opponents.

Figure 5.4 plots the average stated beliefs of each player against their chosen numbers in the game. The observations below the diagonal shows that the large majority of the average beliefs are larger than the chosen numbers. In the first and the fifth rounds, 86% and 72% of adult participants chose numbers which are lower than their average stated beliefs representing more than two-thirds of the sample. The analysis of children revealed a very similar behaviour with 74% and 88% of children selecting numbers lower than their average beliefs in rounds 1 and 5. On the other hand, the convergence of choices towards the Nash equilibrium appears clearer in the sample of adults compared to children. The observations are grouped closely to the bottom left corner in the fifth round with 9% of adults choosing numbers larger than 10 while the corresponding percentage of children was 26%.

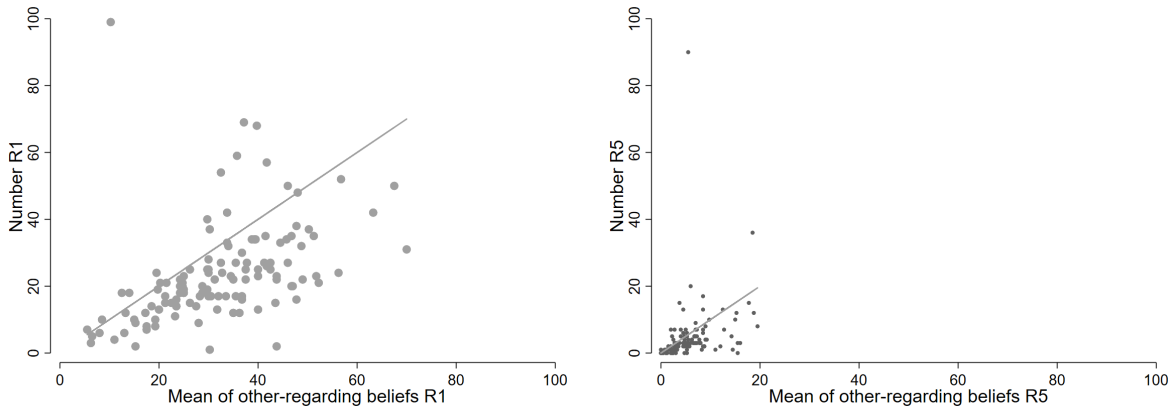

Figure 5.4: Chosen Numbers and Average of Other-Regarding Beliefs

5.2 Depth of Reasoning and Learning

So far, we have shown that choices of adults approach the Nash equilibrium in earlier rounds compared to children. On the other hand, a common finding for both samples is that the average stated beliefs are regularly higher than the chosen numbers. In other words, players seem to adjust their numbers downwards relative to what they previously stated about the other players.

In order to address in more detail the similarities and contrasts between the two samples, we examine the behaviour of adults within the context of the level-k model. The realised levels of reasoning in adults can provide another perspective on their earlier convergence to the Nash equilibrium. Further, considering the elicited levels of reasoning serves to translate the discrepancy between stated beliefs and actions into levels of reasoning. Lastly, we extend the analysis by categorising the behaviour of adults according to the learning direction theory. Throughout this section, we follow the same methods regarding the models and analysis as we described in Section 4.2.

Realised Levels of Reasoning

Table 5.2 displays the relative frequencies of levels of reasoning that adult participants employed in all ten rounds of the Goblin Game. The modal frequencies of each round are reported in bold font. In the first round, only 6% of adults chose dominated strategies ($x_{i1} > 50$), while children that chose numbers above 50 in the first round was three times of that (18%). It can be observed that around 20% of adult players are below the degree zero in rounds 3-5. In other words, in rounds 3, 4 and 5, about one fifth of the players chose numbers above the median number of the previous round. While such behaviour in adults is mostly limited to the middle rounds, it lasted over the ten rounds of the game for children. Interestingly, approximately 20% of children chose numbers that were classified into degree $d < 0$ over all ten rounds while it became more pronounced between rounds 4-6 which corresponds to the middle rounds of the game. Overall, children chose numbers above the median of the previous round more frequently than adults did in the Goblin Game.

In the first three rounds of the game, adult participants mostly employed the degrees of sophistication $d = 1$ and $d = 2$, while we observed most children in the first three rounds

Table 5.2: Relative Frequencies of Levels of Reasoning Over All Rounds

	Round 1	Round 2	Round 3	Round 4	Round 5
d<0	0.06	0.14	**0.20**	**0.21**	**0.22**
d=0	0.07	0.17	0.19	0.15	0.21
d=1	**0.50**	**0.43**	**0.34**	**0.37**	**0.31**
d=2	**0.27**	**0.22**	**0.20**	0.15	0.13
d=3	0.06	0.02	0.04	0.06	0.02
d>3	0.04	0.02	0.03	0.06	0.11
	Round 6	Round 7	Round 8	Round 9	Round 10
d<0	0.16	0.16	0.15	0.10	0.08
d=0	**0.27**	**0.36**	**0.23**	**0.32**	**0.32**
d=1	**0.24**	0.14	0.17	0.11	0.05
d=2	0.19	0.08	0.07	0.02	0.03
d=3	0.01	0.00	0.01	0.00	0.00
d>3	0.13	**0.26**	**0.37**	**0.45**	**0.52**

exhibited $d = 0$ and $d = 1$. That is, most children in the beginning of the game chose numbers either close to the reference point (50 in the first round and the median of the previous round in the rest of the rounds) or p times the reference point. In contrast, adults exhibited a deeper reasoning process due to the fact that so few of them chose numbers close to the reference point. The behaviour of adults in this sense is highly similar to the literature on strategic behaviour in beauty contest games.[26] Thus, our findings suggest that, children exhibit on average shallower degrees of sophistication in the first three rounds of the Goblin Game compared to the sample of adults.

Despite the contrasting findings, the increase in the share of adult players within $d > 3$ points out a common observation for both samples: players employ deeper levels of reasoning through the end of the game. Moreover, as in children, we did not find evidence of increasing depth of reasoning between any consecutive rounds until the seventh round of the game (signtest, p<0.03).

Elicited Levels of Reasoning

In Table 5.3 we report the realised degrees of reasoning of adults together with the elicited degrees of reasoning which is measured with the arithmetic mean of their stated beliefs, for rounds 1 and 5. That is, the degrees of reasoning that each player thinks the others, on

[26]The modal degrees of reasoning Duffy and Nagel (1997) found for the first four rounds of their $1/2 - median$ game was $d = 1$ and $d = 2$.

average, will employ in the next round (mean belief) and the levels of reasoning revealed by players own choices (realised) in the corresponding round. It can be observed that the modal frequencies representing the beliefs about the degrees of sophistication of other players belong to $d = 0$ and $d = 1$ in round 1 and $d < 0$ and $d = 0$ in round 5. In contrast, the numbers chosen by players reveal higher levels of sophistication. We observe some beliefs categorised into higher levels than $d = 1$ contrary to the sharp distribution in the children sample (all children stated beliefs representing that their opponents will exhibit either $d = 1$ or below). Even so, the results provide conformity to the common observation that children and adult players, on average, stated higher numbers for their beliefs than their actions. Indeed, in parallel with children, realised levels of reasoning in the sample of adults are significantly higher than the elicited levels of reasoning, both in round 1 and in round 5 (signtest, $p \approx 0.00$).

Table 5.3: Relative Frequencies of Elicited Levels of Reasoning

Round 1	Realised	Mean Belief
d<0	0.06	0.08
d=0	0.07	**0.33**
d=1	**0.50**	**0.45**
d=2	**0.27**	0.10
d=3	0.06	0.04
d>3	0.04	0.00
Round 5	Realised	Mean Belief
d<0	**0.22**	**0.50**
d=0	0.21	**0.23**
d=1	**0.31**	0.19
d=2	0.13	0.04
d=3	0.02	0.01
d>3	0.11	0.03

Figure 5.5 displays the differences between realised and elicited degrees of sophistication for the adult sample. Similar to children, 43% and %27 of players were only one step ahead of their average beliefs about others, in round 1 and round 5 respectively. In fact, most differences are biased to the right with 57% of them being higher than zero in both rounds.

A Simple Learning Model

Table 5.4 displays the behaviour of adult participants in the Goblin Game classified

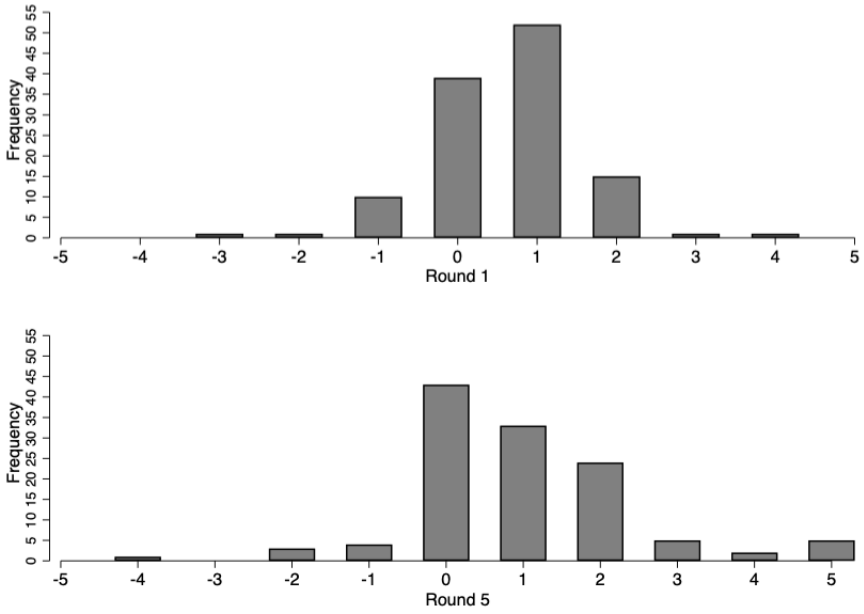

Figure 5.5: Differences Between Realised and Elicited Depths of Reasoning

according to the learning direction theory.[27] Although we found no clear evidence of increasing levels of reasoning in the adults sample, 52% of the choices were, on average, consistent with the learning direction theory in the first three consecutive rounds. Over the ten rounds, on average, 49% of players adjusted their choices in the desired direction after observing the median number in the previous round. The difference of 6 percentage points between children and adults that exhibit consistent behaviour over the ten rounds indicates that more children updated their decisions towards the predicted direction of the learning rule.

5.3 Determinants of Game Performance

Our analysis with the sample of adults showed several similarities and also differences in regard to the behavior of children in the Goblin Game. In this section, we proceed to examine the relevance of cognitive ability and empathy to the game performance. Similar to children, the fluid IQ was measured by the Raven test and empathy by the IRI test as well as the "E on the forehead" task. We also analyse the relationship with the game

[27]Following the same method we described in Section 4.2.3 we exclude the sessions where the median number (and thus the winning number) was zero; the sample sizes are 115, 110, 95, 75 and 65 in rounds 5-6, 6-7, 7-8, 8-9 and 9-10, respectively and full sample (N=120) in the rest of the rounds.

Table 5.4: Relative Frequencies of Behaviour Classified According to the Learning Direction Theory

		Rounds 1-2	**Rounds 2-3**	**Rounds 3-4**
$x > p * median$	Decreased	0.38	0.42	0.49
	Increased	0.50	0.47	0.32
	Constant	0.01	0.00	0.02
$x < p * median$	Increased	0.10	0.07	0.10
	Decreased	0.01	0.01	0.02
	Constant	0.00	0.00	0.01
$x = p * median$	Constant	0.00	0.00	0.00
	Increased	0.00	0.02	0.04
	Decreased	0.00	0.01	0.00
	Consistent	**0.48**	**0.49**	**0.59**
		Rounds 4-5	**Rounds 5-6**	**Rounds 6-7**
$x > p * median$	Decreased	0.38	0.49	0.41
	Increased	0.38	0.28	0.29
	Constant	0.02	0.04	0.15
$x < p * median$	Increased	0.10	0.09	0.05
	Decreased	0.03	0.02	0.01
	Constant	0.02	0.04	0.04
$x = p * median$	Constant	0.00	0.00	0.00
	Increased	0.05	0.04	0.04
	Decreased	0.02	0.00	0.01
	Consistent	**0.48**	**0.58**	**0.46**
		Rounds 7-8	**Rounds 8-9**	**Rounds 9-10**
$x > p * median$	Decreased	0.42	0.31	0.30
	Increased	0.16	0.28	0.18
	Constant	0.15	0.17	0.35
$x < p * median$	Increased	0.10	0.07	0.06
	Decreased	0.01	0.00	0.00
	Constant	0.08	0.08	0.05
$x = p * median$	Constant	0.02	0.03	0.00
	Increased	0.05	0.05	0.00
	Decreased	0.01	0.01	0.06
	Consistent	**0.54**	**0.41**	**0.36**

performance conditional on best responding to stated beliefs considering the accuracy of those statements.

Cognitive Ability and Empathy

Table 5.5 shows the regression estimates of cognitive ability and empathy on the game performance for the sample of adults. In line with children results, the fluid IQ captured by the Raven test in adults does not provide any relevant insights on the game performance. The Goblin Game uses the median, the parameter p equals 1/2 and all possible outcomes were visualised on the game board. The simple design might have suppressed the need for the mental effort in adults compared to a standard beauty contest game, which may explain the insignificant relationship between cognitive ability and game performance. Nevertheless, we can only consider this as a mere possibility given that we do not have enough information to support this claim. On the other hand, empathy skills in adults are not relevant to the game performance contrary to children. Although not significant, the impact of empathy is in the opposite direction and weaker compared to what we observed in children.

Stated Beliefs

Table 5.6 shows the results of the OLS regression with independent variables of best responding to stated beliefs, inaccuracy of stated beliefs and their interactions. In the first round of the game, only 38% of the adults best responded to their stated beliefs, while in round five the portion increased to 52%. These frequencies are not far from that of children (37% and 40%, respectively). However, adults who best responded to their stated beliefs show a significant relationship with the game performance in terms of winning frequency compared to children who exhibit the same behaviour.

On average, adults who best responded to their stated beliefs in the first round have the advantage of almost one extra coin at the end of the game ($p<0.02$). This behaviour remained positively related to the winning frequency as the game progressed. That is, best responding to the stated beliefs in the fifth round leads to an average increment, but this time of almost two coins by the final round (significant at a 1% level). This effect fades away if players are inaccurate to anticipate the decisions of their opponents. Unlike children, none of the interaction terms reveal a significant relationship to the game

Table 5.5: OLS Estimates of Cognitive Ability and Empathy Skills

	(1) Coins R1-R10	(2) Distance to the best response R1-R10
Fluid IQ (Raven)	0.111	0.423
	(0.213)	(0.255)
Empathy (IRI)	-0.064	0.010
	(0.226)	(0.588)
Empathy Task	-0.280	0.607
	(0.430)	(0.649)
Control variables	✓	✓
Session FEs	✓	✓
Observations	113	113
R^2	0.489	0.294

Notes: All independent variables are standardised except for the empathy task which is a dummy variable that takes the value of 1 if the participant drew an "E" on her forehead readable from the interviewer's perspective and 0 otherwise. The vector of control variables includes a dummy variable with value of 1 when the subject i is female and 0 otherwise; the ABITUR; the educational degree; and the age of participants in years squared to account for non-linear relationships with the game performance. There are three missing observations for the empathy task, three for the ABITUR score and one for the educational degree, reducing the sample size to 113 observations. Standard errors clustered at the session level are in parentheses. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

performance. When considered alone, inaccurate beliefs in adults are not strongly related to performance which is similar to the results of children.

Table 5.6: OLS Estimates of Best Responding to Stated Beliefs and Inaccuracy

	(1) Coins R1-R10	(2) Distance to best response R1-R10
Best responding to stated beliefs R1	0.812**	-0.321
	(0.321)	(0.518)
Best responding to stated beliefs R5	1.873***	-0.688
	(0.338)	(0.827)
Inaccuracy R1	-0.027	-0.018
	(0.128)	(0.201)
Inaccuracy R5	0.039	0.275
	(0.120)	(0.310)
Best responding to stated beliefs R1 × Inaccuracy R1	-0.167	0.211
	(0.157)	(0.181)
Best responding to stated beliefs R5 × Inaccuracy R5	-0.032	-0.441
	(0.115)	(0.384)
Control variables	✓	✓
Session FEs	✓	✓
Observations	116	116
R^2	0.694	0.316

Notes: All the independent variables are standardised distances except for *Best responding to stated beliefs* which is a dummy variable. The vector of control variables includes a dummy variable with value of 1 when the subject i is female and 0 otherwise; the ABITUR; the educational degree; and the age of participants in years squared to account for non-linear relationships with the game performance. The ABITUR scores have three missing observations and the educational degree has one, reducing the sample size to 116 observations. Standard errors clustered at the session level are in parentheses. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

To summarise, adults perform more successfully in terms of coins by best responding to their stated beliefs, unless their beliefs about others are increasingly inaccurate. In general for both samples, the ability to correctly guess what others think is not as relevant to the game performance as best responding to stated beliefs. A caveat on these results is that the belief elicitation procedure in adults was rewarded as opposed to children.

6 Conclusion

In this study, we reported the results of a simplified version of the beauty contest game in which players guessed half of the median of all numbers chosen for ten rounds of repetition. The subjects were children between the ages of 8 and 11. Besides collecting information about the choices of players, the game design included a belief elicitation procedure. In addition, the data contained scores on a variety of tests and ratings assessed by the children themselves and also by the instructors who evaluated the behaviour of the players before and during the game. The choices of children in the first round of the game is consistent with the literature on beauty contest games. The distribution of children's actions over rounds present few differences with a sample of adults playing the same game; at the beginning of the game, children tend to choose dominated strategies more frequently. We also demonstrate in many ways that children's choices converge towards the Nash equilibrium, although in later rounds and with a wider dispersion compared to adults. By the end of the game, 40% of children choose zero whereas 53% of adults play the Nash equilibrium in the final round.

We also looked into the degrees of sophistication revealed by the choices of children as well as by their stated beliefs. Using the model of depth of reasoning, we find further evidence that the behaviour of children was not entirely random. However, the modal degree being $d = 1$ indicated that most children choose a best response to the random behaviour of $d = 0$ (Nagel, 1995). In the last rounds of the game, the levels of reasoning classified into $d > 3$ representing the eventual convergence of choices towards the Nash equilibrium. In the game for adults, modal frequencies of strategic sophistication indicated the levels $d = 1$ and $d = 2$ in the first three rounds of the game. This result suggests that adults were able to reason one step deeper than children in the initial rounds of the game. Besides, the degree $d > 3$ reached the highest concentration of observations one round earlier than children in the sample of adults. Moreover, children as well as adults have revealed levels of sophistication in terms of their choices that are on average higher than their elicited levels of sophistication. This finding is compatible with the theory of the level-k model in which players act with a deeper level of reasoning than they expect from their opponents. The application of the learning direction theory allowed us to explore the

adjustment of choices from one round to the next in accordance with the learning model. Children were consistent to the model's predictions at an average frequency of 54% for all rounds. In a similar way, the portion of adults whose behaviour fits the mentioned learning model was 49% on average throughout the game.

Once we provide evidence to claim that our data is relevant to study children's behaviour in the beauty contest game, we took the analysis forward and investigated how players' individual characteristics are related to the performance in our variant of the beauty contest game. First, we found that children's ratings of understanding are not as reliable as those of external observers. Interviewers were able to identify players with good understanding of the game and whose behaviour was related to successful performance in the upcoming rounds. In addition, empathy skills of children strongly relates to a higher winning frequency which does not happen in the sample of adults. Finally, we analyse how successful were those subjects who best responded to their stated beliefs, and tested if the relationship varies depending on the accuracy of their beliefs. For both samples, children and adults, best responding to inaccurate stated beliefs is negatively related to the game performance. Nevertheless, the estimates are statically significant only for children and for the measure of performance based on the winning frequency over the rounds. The choices of children are less often a best response to their own stated beliefs (37% in round 1 and 40% in round 5) compared to adults (38% and 52%).

The conclusions we infer from the complementary analysis with the sample of adults should be considered with some caveats. Before the Goblin Game, adults played a regular beauty contest game in the form of "guess 2/3 of the average", which might have altered their performance afterwards. Also, we cannot discard potential differences that may arise due to a rewarded belief elicitation procedure on the accuracy of the stated beliefs. Besides, whether the elicited beliefs are true representatives of the underlying beliefs remains as a general limitation for both samples.

The impact of age upon the development of rational behaviour and strategic thinking in children has been documented broadly in the literature. The differences we observe between our samples can suggestively be attributed to the effect of age on the performance in the beauty contest game. However, we cannot provide reasons for the observed differences in the behaviour of children relative to adults. Such an analysis exceeds the scope of our

research questions. Future research can shed light on this question with insights from psychology. To the best of our knowledge, the Goblin Game was the first study to adapt the beauty contest game to children with a belief elicitation procedure. Now that there is evidence suggesting that children exhibit a significant behaviour, there are prospects to conduct more studies on how individuals of various age groups behave and what is the role of their beliefs in beauty contest games.

References

Agranov, M., Caplin, A., and Tergiman, C. (2013). Naive play and the process of choice in guessing games. *Journal of the Economics Science Association*, pages 1–12.

Apesteguia, J., Huck, S., Oechssler, J., Weidenholzer, E., and Weidenholzer, S. (2018). Imitation of peers in children and adults. *Games*, 9(1):11.

Barash, J., Brocas, I., Carrillo, J. D., and Kodaverdian, N. (2019). Heuristic to bayesian: The evolution of reasoning from childhood to adulthood. *Journal of Economic Behavior & Organization*, 159:305–322.

Bosch-Domenech, A., Montalvo, J. G., Nagel, R., and Satorra, A. (2002). One, two,(three), infinity,...: Newspaper and lab beauty-contest experiments. *American Economic Review*, 92(5):1687–1701.

Bosch-Domènech, A., Montalvo, J. G., Nagel, R., and Satorra, A. (2010). A finite mixture analysis of beauty-contest data using generalized beta distributions. *Experimental economics*, 13(4):461–475.

Brañas-Garza, P., Garcia-Munoz, T., and González, R. H. (2012). Cognitive effort in the beauty contest game. *Journal of Economic Behavior & Organization*, 83(2):254–260.

Breitmoser, Y. (2012). Strategic reasoning in p-beauty contests. *Games and Economic Behavior*, 75(2):555–569.

Brocas, I. and Carrillo, J. D. (2018a). The determinants of strategic thinking in preschool children. *PloS one*, 13(5):e0195456.

Brocas, I. and Carrillo, J. D. (2018b). Iterative dominance in young children: Experimental evidence in simple two-person games. *Journal of Economic Behavior & Organization*.

Brosig-Koch, J., Heinrich, T., and Helbach, C. (2015). Exploring the capability to reason backwards: An experimental study with children, adolescents, and young adults. *European Economic Review*, 74:286–302.

Burfurd, I. and Wilkening, T. (2018). Experimental guidance for eliciting beliefs with the stochastic becker–degroot–marschak mechanism. *Journal of the Economic Science Association*, 4(1):15–28.

Burnham, T. C., Cesarini, D., Johannesson, M., Lichtenstein, P., and Wallace, B. (2009). Higher cognitive ability is associated with lower entries in a p-beauty contest. *Journal of Economic Behavior & Organization*, 72(1):171–175.

Camerer, C. and Ho, T.-H. (1998). Experience-weighted attraction learning in coordination games: Probability rules, heterogeneity, and time-variation. *Journal of mathematical psychology*, 42(2-3):305–326.

Camerer, C. F. (2010). Behavioural game theory. In *Behavioural and Experimental Economics*, pages 42–50. Springer.

Camerer, C. F., Ho, T., and Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898.

Camerer, C. F., Ho, T.-H., Chong, J.-K., Weigelt, K., et al. (2002). Strategic teaching and equilibrium models of repeated trust and entry games. *Unpublished Caltech manuscript, October*.

Chou, E., McConnell, M., Nagel, R., and Plott, C. R. (2009). The control of game form recognition in experiments: Understanding dominant strategy failures in a simple two person "guessing" game. *Experimental Economics*, 12(2):159–179.

Coricelli, G. and Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, 106(23):9163–9168.

Costa-Gomes, M. A. and Weizsäcker, G. (2008). Stated beliefs and play in normal-form games. *The Review of Economic Studies*, 75(3):729–762.

Czermak, S., Feri, F., Glätzle-Rützler, D., and Sutter, M. (2016). How strategic are children and adolescents? experimental evidence from normal-form games. *Journal of Economic Behavior & Organization*, 128:265–285.

Danz, D. N., Fehr, D., and Kübler, D. (2012). Information and beliefs in a repeated normal-form game. *Experimental Economics*, 15(4):622–640.

Duffy, J. and Nagel, R. (1997). On the robustness of behaviour in experimental 'beauty contest'games. *The Economic Journal*, 107(445):1684–1700.

Fehr, E., Bernhard, H., and Rockenbach, B. (2008). Egalitarianism in young children. *Nature*, 454(7208):1079.

Gilet, A.-L., Mella, N., Studer, J., Grühn, D., and Labouvie-Vief, G. (2013). Assessing dispositional empathy in adults: A french validation of the interpersonal reactivity index (iri). *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 45(1):42.

Gill, D. and Prowse, V. (2016). Cognitive ability, character skills, and learning to play equilibrium: A level-k analysis. *Journal of Political Economy*, 124(6):1619–1676.

Grosskopf, B. and Nagel, R. (2007). Rational reasoning or adaptive behavior? evidence from two-person beauty contest games. *Evidence from Two-Person Beauty Contest Games (June 2007). Harvard NOM Research Paper*, (01-09).

Harbaugh, W. T., Krause, K., and Berry, T. R. (2001). Garp for kids: On the development of rational choice behavior. *American Economic Review*, 91(5):1539–1545.

Hermes, H., Hett, F., Mechtel, M., Schmidt, F., Schunk, D., and Wagner, V. (2019). Do children cooperate conditionally? adapting the strategy method for first-graders. *Journal of Economic Behavior & Organization*.

Hyndman, K., Özbay, E. Y., Schotter, A., and Ehrblatt, W. (2012). Belief formation: an experiment with outside observers. *Experimental Economics*, 15(1):176–203.

Hyndman, K. B., Terracol, A., and Vaksmann, J. (2013). Beliefs and (in) stability in normal-form games. *Available at SSRN 2270497*.

Keynes, J. M. (1936). The general theory of employment, interest and money.

Khadjavi, M. and Nicklisch, A. (2018). Parents' ambitions and children's competitiveness. *Journal of Economic Psychology*, 67:87–102.

Kocher, M., Sutter, M., and Wakolbinger, F. (2014). Social learning in beauty-contest games. *Southern Economic Journal*, 80(3):586–613.

Kocher, M. G. and Sutter, M. (2004). The decision maker matters: Individual versus group behaviour in experimental beauty-contest games. *The Economic Journal*, 115(500):200–223.

Lahav, Y. (2015). Eliciting beliefs in beauty contest experiments. *Economics Letters*, 137:45–49.

Manski, C. F. and Neri, C. (2013). First-and second-order subjective expectations in strategic decision-making: Experimental evidence. *Games and Economic Behavior*, 81:232–254.

Meindl, C. (1998). Feas-fragebogen zur erfassung von empathie und angemessenem sozialem verhalten. *ZPID (Leibniz Institute for Psychology Information)*.

Moreira, B., Matsushita, R., and Da Silva, S. (2010). Risk seeking behavior of preschool children in a gambling task. *Journal of Economic Psychology*, 31(5):794–801.

Moulin, H. (1986). *Game theory for the social sciences*. NYU press.

Müller, J. and Schwieren, C. (2011). More than meets the eye: an eye-tracking experiment on the beauty contest game. Technical report, Discussion Paper Series.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review*, 85(5):1313–1326.

Nagel, R., Bühren, C., and Frank, B. (2017). Inspired and inspiring: Hervé moulin and the discovery of the beauty contest game. *Mathematical Social Sciences*, 90:191–207.

Nyarko, Y. and Schotter, A. (2002). An experimental study of belief learning using elicited beliefs. *Econometrica*, 70(3):971–1005.

Sbriglia, P. (2008). Revealing the depth of reasoning in p-beauty contest games. *Experimental Economics*, 11(2):107–121.

Schotter, A. and Trevino, I. (2014). Belief elicitation in the laboratory. *Annu. Rev. Econ.*, 6(1):103–128.

Selten, R. and Stoecker, R. (1986). End behavior in sequences of finite prisoner's dilemma supergames a learning theory approach. *Journal of Economic Behavior & Organization*, 7(1):47–70.

Stahl, D. O. (1996). Boundedly rational rule learning in a guessing game. *Games and Economic Behavior*, 16(2):303–330.

Sutter, M., Czermak, S., and Feri, F. (2013). Strategic sophistication of individuals and teams. experimental evidence. *European economic review*, 64:395–410.

Sutter, M., Zoller, C., and Glätzle-Rützler, D. (2019). Economic behavior of children and adolescents–a first survey of experimental economics results. *European Economic Review*, 111:98–121.

Appendices

A Distribution of choices in the Goblin Game

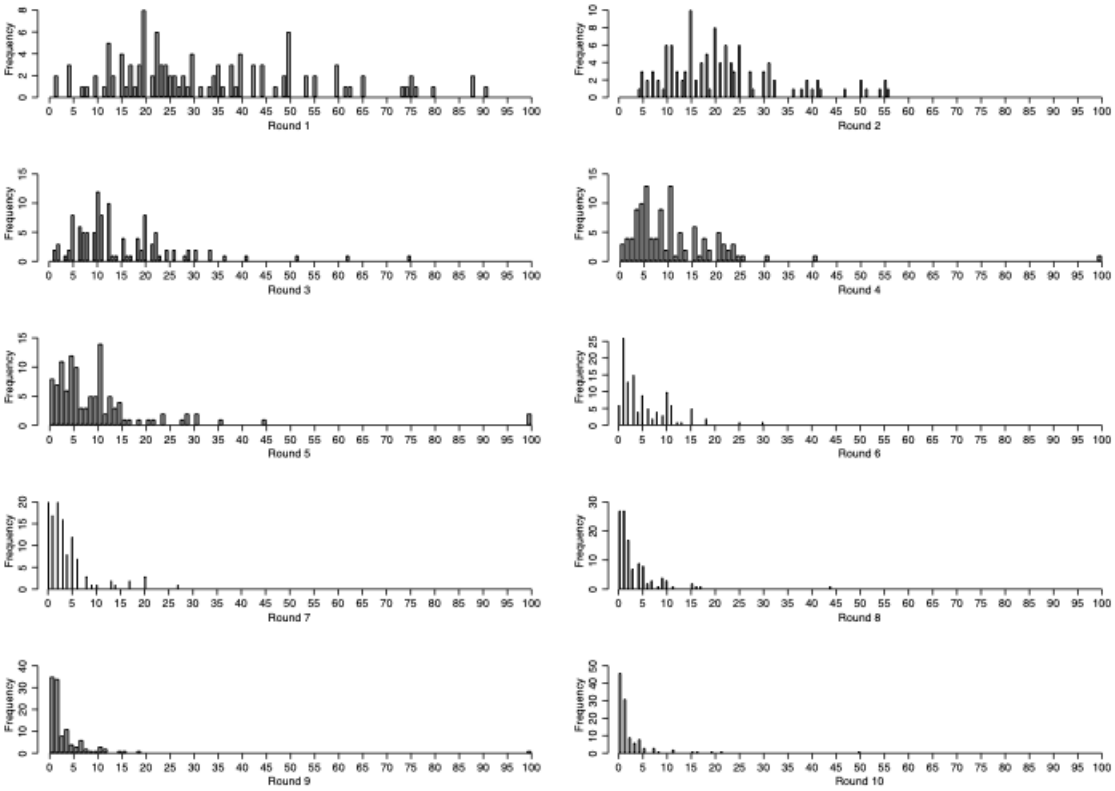

Figure 6.1: Children: Histograms of Chosen Numbers in Each Round

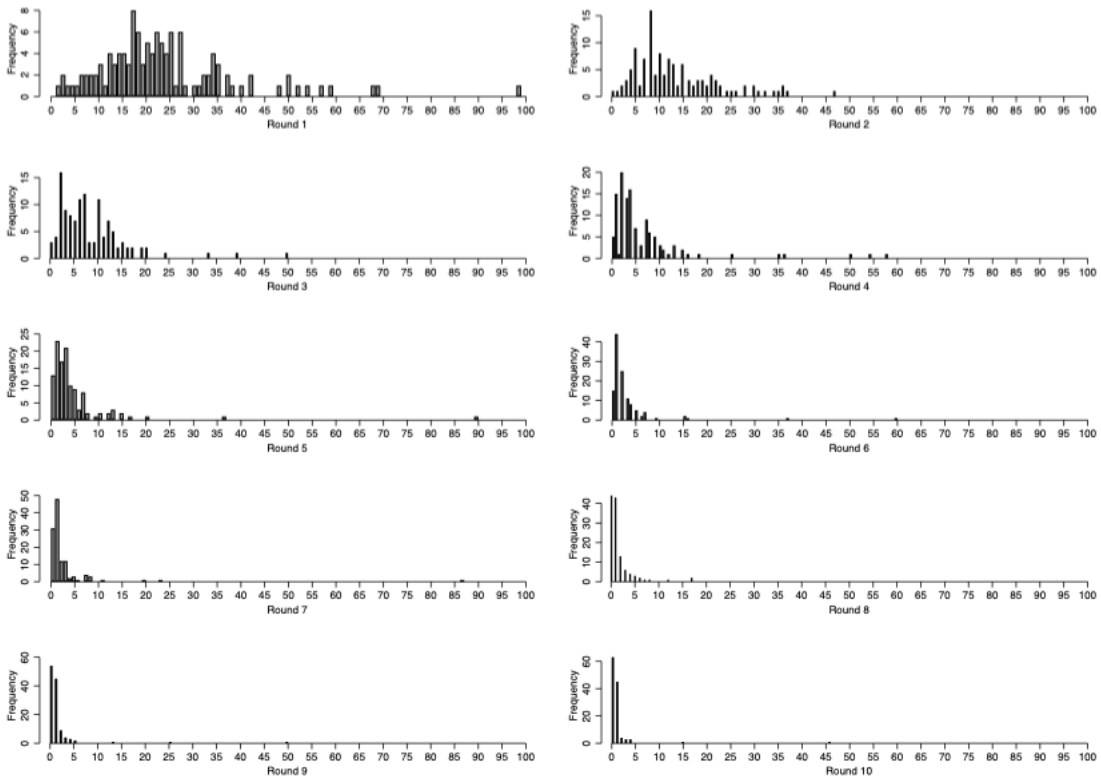

Figure 6.2: Adults: Histograms of Chosen Numbers in Each Round

B Self-regarding beliefs

Figure 6.3 plots stated beliefs that children inform about themselves and the corresponding actual numbers they played in rounds 1 and 5 of the game. The shape of a 45-degree line, formed by the matched observations in both rounds, indicates that some players stick to their self-regarding beliefs about the number to choose in the game. On the other hand, the dispersion of the observations posits otherwise. Indeed, 42 and 65 of the players actually chose the same number as they stated in the first round and in the fifth round, respectively. We interpret the behaviour of children towards the self-regarding beliefs as ambiguous.

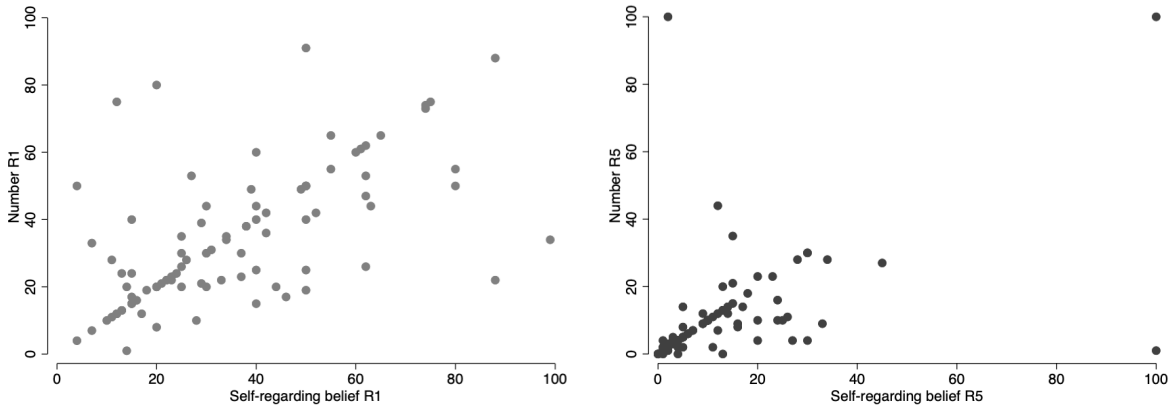

Figure 6.3: Chosen Numbers and Self-Regarding Beliefs