Experimental Economics (2022) 25:229–253
https://doi.org/10.1007/s10683-021-09713-y

ORIGINAL PAPER

If you could read my mind–an experimental beauty-contest game with children

Henning Hermes[1] · **Daniel Schunk**[2]

Received: 6 January 2020 / Revised: 16 March 2021 / Accepted: 23 March 2021 / Published online: 4 May 2021

Abstract

We develop a new design for the experimental beauty-contest game (BCG) that is suitable for children in school age and test it with 114 schoolchildren aged 9–11 years as well as with adults. In addition, we collect a measure for cognitive skills to link these abilities with successful performance in the game. Results demonstrate that children can successfully understand and play a BCG. Choices start at a slightly higher level than those of adults but learning over time and depth of reasoning are largely comparable with the results of studies run with adults. Cognitive skills, measured as fluid IQ, are predictive only of whether children choose weakly dominated strategies but are neither associated with lower choices in the first round nor with successful performance in the BCG. In the implementation of our new design of the BCG with adults we find results largely in line with behavior in the classical BCG. Our new design for the experimental BCG allows to study the development of strategic interaction skills starting already in school age.

Keywords Children · Experimental beauty-contest game · Guessing game · Strategic interaction · Decision-making · Cognitive skills · Noncognitive skills

JEL Classification C72 · C91 · D91

✉ Henning Hermes
henning.hermes@nhh.no

[1] FAIR/Department of Economics, NHH Bergen, Helleveien 30, 5045 Bergen, Norway

[2] Department of Economics, Johannes Gutenberg University of Mainz, Jakob-Welder-Weg 4, Mainz 55128, Germany

1 Introduction

An important skill for economic actors is the ability to anticipate the actions of others in strategic settings and to choose one's own actions accordingly. There is a very wide range of situations in which these strategic interaction skills are crucial.[1] For example, in their well-known study on signaling games, Cooper and Kagel (2005) argue, based on recording and coding of the dialogues of experimental participants, that "a critical step in monopolists' learning to play strategically is putting themselves in the entrant's shoes, reasoning from the entrant's point of view to infer likely responses to their choice as a monopolist." Similarly, empirical evidence demonstrates that individual investors base their investment decisions on their beliefs about the return expectations of *other* investors (Rangvid et al., 2013; Egan et al., 2014). Another example are most forms of matching markets; specifically, Braun et al. (2014) show this in the context of a laboratory study that investigates students' behavior in university admissions procedures. Finally, the dynamics and outcomes of weakest-link games, representing, for example, coordination problems in a firm, depend on the players' ability to anticipate the actions of their peers (e.g., Brandts and Cooper, 2006).

But why studying these skills in children? In the last two decades, a large number of studies (summarized in Kautz et al., 2014) has documented that preferences and skills are shaped in the early years, they form the basis for future investments, and fundamentally determine adult life outcomes. For this reason, the early development of preferences and skills has been studied in great detail in recent years in the economic literature (see Sutter et al., 2019, for a review). By contrast, children's strategic interaction skills have been studied much less, despite being of similar importance in their own right as well as for the life cycle of skills. The study of strategic interaction skills among children is inherently difficult, reflecting Jean Piaget's view that up to a certain age, children systematically lack the ability to adopt other's perspectives (Piaget, 1962). Hence, most existing work on this topic has so far considered only settings, in which children interact either with a single peer child or with a computer (see the literature review below). However, this disregards the fact that most strategic interactions outside the laboratory involve several (and often large numbers of) peers, and thus require a more comprehensive notion of strategic sophistication.

Here, we develop a novel design to study strategic interaction skills in children, which is based on the experimental beauty-contest game (henceforth, BCG) introduced by Nagel (1995, at that point called "guessing game"): N decision-makers simultaneously choose a number x between 0 and 100 and the person with the number closest to $p * \bar{x}$ wins a fixed prize (mostly, $p = 2/3$). This economic game has been used to study strategic interaction in groups for more than two decades. The experimental BCG has a strong advantage compared with most other interactive games used in economics to study decision-making (e.g., ultimatum game, public

[1] Brocas and Carrillo (2018b) define strategic interaction to be the intrinsic ability to anticipate the actions of others and to act accordingly.

goods game, etc.): Neither social, nor time or risk preferences should affect decision-making (cf. Kocher & Sutter, 2005).[2] A substantial body of knowledge has been developed concerning how various game parameters such as repetition, feedback, or time pressure affect performance in a BCG (Duffy & Nagel, 1997; Ho et al., 1998; Weber, 2003; Kocher & Sutter, 2006) and how individuals versus teams of different sizes behave (Kocher & Sutter, 2005; Sutter, 2005). Yet, there is no study using a BCG to measure strategic interaction skills in groups with children. Given the relevance of strategic interaction skills for economic decision-making (see above), having such a measure of strategic interaction skills for children would not only be interesting in itself but would also enable us to shed light on which other skills (cognitive or noncognitive) are the building blocks to developing strong strategic interaction skills over the life cycle.

To achieve this goal, we simplify the experimental BCG into a board game—we make it less abstract, provide concrete and visually illustrated operations with a spatial interpretation corresponding to each step in the game, and use the median, integers (0–100), $p = 1/2$, and only five players per group. Applying this new design for the experimental BCG, we study the behavior of $n = 114$ children aged 9–11 years in the game and demonstrate that they are capable of successfully playing an experimental BCG. We also validate the new design using an adult student sample with $n = 120$. In a second step, aimed at understanding the building blocks of children's strategic interaction skills, we analyze the link between children's performance in this game and their fluid IQ, an important part of cognitive skills. Our results demonstrate that fluid IQ is predictive only of whether children choose weakly dominated strategies, but is not associated with lower choices in the first round or with successful performance in the subsequent BCG. Hence, we contribute to the literature on strategic interaction in two ways, namely by (1) developing a tool to study strategic interaction in groups with young children and adolescents, and by (2) analyzing the role of cognitive skills for strategic interaction in a less abstract game form.

Our study is related to a literature on strategic interaction in children. Some early studies (starting with Murnighan & Saxon, 1998; Harbaugh & Krause, 2000; Harbaugh et al., 2003a, 2003b) have examined children's behavior in simple interactive games, such as dictator, ultimatum, trust, and public goods games. However, there is not much research about more complex strategic interactions in children, i.e., settings in which the cognitive demand of the task is higher and in which it matters to what extent children are able to take the perspective of others and translate this into their own strategic decision-making.[3]

[2] To keep instructions simple, in our design children receive a payoff in *every* round of the BCG (see Sect. 2.3 for details). In this setting, social preferences actually could affect decision making, for example, because children might be inequality averse.

[3] We should mention that there is a literature in developmental psychology that studies the extent to which children are able to take the perspectives of others (see Birch et al., 2017, for a summary). This literature focuses on how children reason about other children's motives and decisions, but not on the strategic interactions of the children.

Among the few notable exceptions are the following studies: First, there is a study on strategic interaction that involves observing child-experimenter interactions: Sher et al. (2014) have children play two games, a sticker game, and a sender-receiver game. They argue that for successful strategic interaction, children require the ability to undertake recursive thinking, i.e., "the ability to use the output of one step of a reasoning process as input to a following step" as well as the ability to put themselves into another person's shoes, i.e., to understand another person's motives and emotions ("Theory of Mind") as well as their incentives (again, cf. Cooper & Kagel, 2005). In their study, many children possess these skills from about the age of 7 years onward. Second, there is a number of two-person games, i.e., involving the interactions between two children: Brocas and Carrillo (2018a) examine iterative reasoning among children from pre-kindergarten to first grade in the context of four two-person games. They show that both logical thinking and Theory of Mind are key ingredients to successful strategic reasoning, and trace children's limitations in both these skills in their age groups. In another series of two-person games, Brocas and Carrillo (2018b) present further support for these findings and show that although preschoolers are able to think strategically in principle, this does not mean that they are capable of acting accordingly. Brocas and Carrillo (2020) play a two-person BCG with children and adolescents from 5 to 18 years. Their findings indicate that strategic sophistication is present early on and, while abstract reasoning is known to develop through adolescence, equilibrium play in their game improves through childhood and stabilizes already after age 10. The detected learning trajectory is mostly due to feedback-learning, i.e., observing the choice of the partner. Fe et al. (2019) play a competitive game designed to trigger level-k thinking with pairs of children aged between 5 and 12 and find that Theory of Mind, cognitive ability, and age are related to the children's level-k-behavior. Furthermore, Czermak et al. (2016) present 10–17-year-old children with simple experimental normal-form games for two children and elicit their beliefs, providing evidence that children of this age are clearly able to strategize in their choices. Third, there is one study that involves the interaction between three children: Brocas and Carrillo (2019) use a graphical paradigm of a dominance solvable game to study the evolution of the ability to think strategically between 8 years and adulthood. They report significant performance increases between 8 and 12 years and a stabilization afterwards.

While these above-mentioned studies are related to our approach, our experimental setting involves the interaction with a *group* of other children in the context of a normal multi-player BCG. That is, the game requires a substantive degree of logical thinking and perspective-taking. Research by Brosig-Koch et al. (2015) has demonstrated that children at the age of 6 years already have the ability to reason backwards and that this ability increases with age. Thus, we expect that children in our sample (aged 9–11 years) will possess the necessary cognitive skills.[4] Moreover,

[4] Castillo et al. (2018) show with a sample of 7th–9th graders that although many observed choices are not completely consistent with predictions from economic theory, they are not random either. They interpret this as an indication of rational behavior, which further supports that children of our age group should have the necessary level of rational thinking.

empirical research using Theory of Mind tasks confirms that, by the end of preschool, most children should be well able to take the perspective of others (Wellman et al., 2001; Brocas & Carrillo, 2018a); hence, we expect children to also possess sufficient perspective-taking abilities. Nevertheless, strategic interaction in the form of an experimental BCG is very challenging, even for adults. Grosskopf and Nagel (2008) conducted a two-player version of the BCG, which can be easily translated into a "whoever chooses the smaller number wins" contest. They found that only 10% of (adult) lab participants and 37% of professionals (i.e., participants at economic conferences) actually chose zero—in this form of the game, zero is the dominant strategy, irrespective of the other player's choice (means were 35.6 and 21.7, respectively). Similarly, Bosch-Rosa and Meissner (2020) use one player guessing games with adults to show that even with this simple structure, many subjects fail to fully understand the structure of the game. The findings of our study suggest that—despite this challenging setting—children are able to understand the experimental BCG if it is presented in a visual manner, and we use this new design to draw conclusions on the determinants of strategic interaction behavior in children.

The remainder of this paper is structured as follows: Sect. 2 describes the experimental procedures and the new design of the experimental BCG used in the present study, Sect. 3 presents the results and discussion, and Sect. 4 concludes.

2 Experimental design

The study was conducted in schools with children aged 9–11 years (third and fifth grade). The following subsections provide details on the procedures, participants, the new design of the BCG, and the outcomes measured.

2.1 Procedures

We conducted the study in March and April 2016 in three different schools in Germany. We contacted the three schools, asked for their participation (all agreed), and sent parental consent forms for data use to the teachers of the participating classes (we were not involved in the selection of classes). Six classes were tested, each within one day. Days were always structured in the same way. In the first lesson (45 min), the experimenter informed the children about the procedures of the day, including the incentives they could win by earning "gold coins". Then, the experimenter guided the class through a workbook to collect data on a number of control outcomes using various questionnaires and tests. All questionnaires were read out loud to avoid problems with different levels of reading skills. In the subsequent lessons, we took randomly formed groups of five children out of the class and played the BCG with them in a separate room (again, each lesson lasted about 45 min). Groups were formed based on numbers on children's workbooks which were randomly distributed in the classroom in the first lesson. After the last lesson, children

went to a separate room, where they could trade the gold coins they won during the game for toys.

2.2 Participants

Six classes in three different schools participated in the study; four classes were in the third grade (in total 77 children), two were in the fifth grade (in total 50 children). Overall, 113 of these children plus two children from another class played the BCG (a total of 115 children). The parents of one of these children did not provide consent for data use, resulting in a final sample of 114 children.[5] Children played the BCG in 23 groups of five children. Each group consisted of randomly selected children from one single class, with two exceptions: in one case, a child from class A played together with four children from class B because this child had not been able to play the BCG the day before due to time constraints. In the other case, we recruited two children from another class to ensure that we had a complete group of five children playing the BCG. In the final sample, 63 children were female (55%), 64 were in the third grade (56%), and the mean age was 10 years (SD = 1, see Table A1 for details).

2.3 The beauty-contest game

At the heart of the study was an incentivized experimental BCG in a new design. To facilitate understanding of the game, we (1) used a board game setup, so children had a visualization as well as a spatial mapping of the game in front of them, (2) embedded the rules in a story-like design that was easy to recall and very simply structured, and (3) ensured the highest possible degree of understanding by explaining the game in a one-to-one setting between each child and a trained experimenter, with each child asked to explain each step of the game back to the experimenter. We also simplified the game parameters to make the game easier to understand. Specifically, we used $p = 1/2$, the median (instead of the mean), five players (a group size for which strategic environment effects are triggered, see Hanaki et al., 2019), and integers in the range from 0 to 100.

The Board Game. We developed a board game mirroring the experimental BCG called the "Goblin Game" (see Fig. 1). The board depicts the range of integers from 0 to 100 (in black and white) as the range of numbers from which to choose, and a second range of numbers (in green) from 0 to 100 in steps of 1/2 to indicate the winner. In the center of the board, there was an open treasure chest containing gold coins (the prize that children could win in each round of the game). At zero, the goblin had his starting position (as he would always start from zero and would then

[5] We make no claim about the representativeness of our sample. Although we achieve a high participation rate (113 out of 127 children), it is possible that there is selective participation on the individual and/or the school level.

"walk up" to the median player). Numbers were arranged in a circle to create an illusion of "equality of numbers", i.e., there was no "best" or "worst" number.

Children played the BCG at a large table in a separate room exclusively used for the study on the day of data collection. Each child was randomly assigned to one of the colors—yellow, blue, orange, white, or gray—and received a game piece (henceforth called a "pawn") in their respective color.[6] The color also determined each child's seating position at the table. One experimenter sat at the head of the table and led the game (see Figures A4–A6 in the Appendix for details on the experimental setup and the material used in the game).

The Rules of the Game. In order to make the experimental BCG as easy to understand as possible, we structured the rules of the game into a five-step procedure that was repeated for every round played. Each group played 10 rounds. The five steps within each round were as follows.

1. All players secretly write down a number between 0 and 100 on their game slip.
2. All players simultaneously place their pawn onto the number from step 1 on the board.
3. The goblin starts from zero and walks up to the third, i.e., the "middle" player.
4. Having reached the third player, the goblin jumps back by half the distance he has walked.
5. The player who is now closest to the goblin wins a gold coin.

No communication between children was allowed during the 10 rounds of the game. As practical measures, we told children to indicate that they had made their decision in step 1 by covering the number they had written on their game slip with their pawn. In this way, we avoided children writing or changing numbers after having seen choices of their peers. Of course, we also emphasized (and ensured) that in step 2 children had to put their pawn on the actual number they wrote down on their game slip. To help children avoid confusion between the numbers they chose themselves and the numbers on which the goblin was operating, we color-coded the two arrays of numbers: children used the larger, black-and-white numbers, whereas the (green) goblin used the green, inner circle of numbers. For children having problems calculating what "jumping back by half" meant, we printed the number that the goblin would jump to on each of the green fields (e.g., the green field for 30 would read "30 → 15"). For simplicity and because we could not split the gold coins, if several players were equally close to the goblin, each of them would receive a coin.[7]

At the same time, we embedded the rules in a story-like design. Children learned that the goblin had "hidden gold coins in the forest" but that he would be willing to help children find the coins and would "reveal the location of the gold coins to the middle player". Also, children were told that the goblin was "hexed"—for each step he would go forward, he would have to take half a step backward. At the end of each

[6] We tried to avoid using colors with strong meanings, such as red, green, pink, etc.

[7] We should point out that, in principle, this game feature could imply an incentive for children to coordinate. However, the reason for this design choice was to keep the game design as simple as possible.

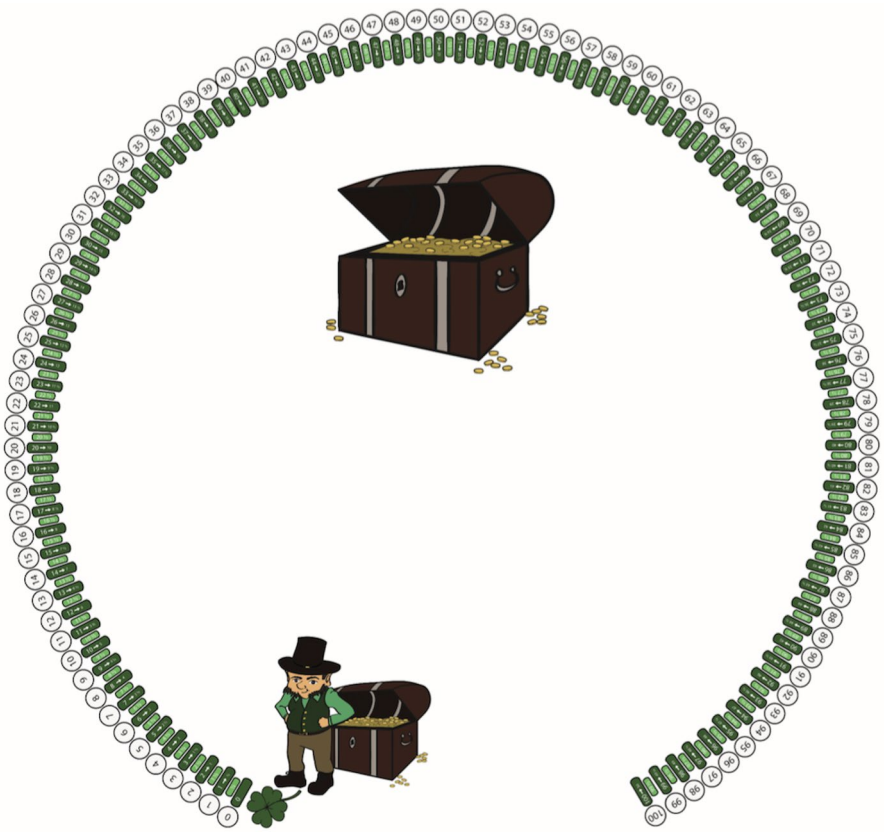

Fig. 1 Design of the Beauty-Contest Board Game ("Goblin Game")

round, the goblin would give a gold coin to the player who was closest to him (see Sect. F.2 in the Appendix for the exact instructions).

Explaining the Rules. As this study was the first involving an experimental BCG played with children, a major challenge was to ensure that the game was properly understood by all children. Consequently, we had five trained experimenters present in each group to (1) ensure the best opportunities for each child to understand the game and (2) to check the individual understanding of each child. Hence, every child individually received a (standardized) explanation from an experimenter and could ask the experimenter any questions that he or she had regarding the game. In this part of the lesson, each child sat at an individual small table distributed around the classroom (cf. Figure A6 in the Appendix). To explain the game, the experimenter used a written instruction, which he/she read to the child, a small version of the board game, and the relevant materials (coins and pawns, see Figure A5 in the Appendix for the setup of an individual table).

At the beginning of the explanation, children were told that they would now play a game in which they could earn gold coins. These coins could later be exchanged

for a real toy and the more coins they earned, the larger would be the selection of toys from which they could choose. Children were encouraged to "try hard" to earn as many coins as possible. By starting the game this way, we wanted to (1) motivate the children and make incentives salient, and (2) make sure that the children understood that they could actually "do something" about the number of coins won and that the game was by no means pure luck. Subsequently, the rules of the game were explained to the children. Then, they could ask questions in case anything was not clear. Finally, the experimenter asked each child to explain back to him/her each of the five steps of the game, one by one. For each step, the experimenter rated the child's understanding (see Sect. 2.4). Afterwards, the experimenter expressed appreciation for the child's explanations and "awarded" him or her a first gold coin (this was done to avoid children having zero coins while playing the game). At the end of the individual explanation session, the experimenter conducted a very short task with the child (see Sect. D in the Appendix) and subsequently sent him or her to the large table in the middle of the classroom to play the "Goblin Game" with his or her classmates.

All research assistants acting as experimenters participated in a half-day training session, which included information on general procedures, a discussion of the importance of standardized instructions, how to deal with potential questions by children, and how to organize the data collection. General hints for experimenters included not reacting to strategies or suggestions articulated by children on how to win the game, and to answer questions only by referring to the general rules of the game. Experimenters were explicitly instructed not to mention any potential strategies to the children. The exact instructions read out to each child can be found in Sect. F.2 in the Appendix.

Playing the Game. Once all children had finished their instructions with their individual experimenter, they sat down at the large table in the middle of the classroom (see Figure A4). Each seat was clearly marked with the color of each pawn; hence, seating positions were exogenous. Before starting the first round, the main experimenter once again repeated the five steps of the game. He then guided the children through all 10 rounds of the game, each time structured by the five steps. At the end of each round, the experimenter gave a gold coin to the winning child. After the last round, individual workbooks and game slips (containing the numbers for each round) were collected and each child received another gold coin for his or her workbook. To avoid any communication about the game between children who already played the game with children who had not yet played the game, we then explicitly instructed children not to tell their peers anything about the game. Finally, children were brought back to their classroom and the next group was picked up to play the BCG in a separate room.

2.4 Outcomes and data

To account for children's choices and performance during the BCG, we collected the individual workbooks and game slips filled out during the game. In addition, we collected the experimenters' ratings on children's understanding of the game (see

Sect. 2.3). To analyze to what extent children's cognitive skills are determinants of successful performance in strategic interaction games, we elicited children's fluid IQ. We used a well-established measure feasible for children, namely Raven's "Coloured Progressive Matrices" (Bulheller & Häcker, 2010). Based on the collected data, we constructed the following outcome and control variables:

Numbers chosen. We used the numbers each child wrote down on his or her individual game slip during the BCG.[8]

Number of Coins Won. A straightforward measure of successful performance in the game is the number of coins each child won during the game. This is equivalent to the number of rounds (out of 10) that a child won. Note that in the case of a tie, more than one child could win a coin. Thus, the total number of coins per group could be larger than 10.

Distance to Best Response. The number of coins won may not fully reflect the degree to which a child made successful choices during the game. For instance, a child could miss out on winning by only one step, whereas another child could miss out on winning by 30 steps; yet, they would be equal with respect to the number of coins they won. Therefore, to analyze successful choices beyond the simple indicator of *winning* a round, we constructed the following outcome measure. We computed the best response for each player in each round (i.e., given the other players' choices in this round, the number that would be the ideal response to win this round—this number is equivalent to half of the second-smallest number of the other four players). We then calculate the distance between the player's actual choice and this best response, take the absolute value, and divide the distance by half the median number in this group in this round to account for scaling differences over the course of 10 rounds (if the median number was zero, we divide by one). Therefore, for each player, the measure of distance reports how far away (measured in fractions of half the median in this round) a player was from his or her ideal response to the other four players' choices.

Rank Based on Distance. Because the "distance to best response" measure is difficult to compare across (1) groups and (2) rounds (despite the normalization with half the median in each round), we rank players within each group within each round based on their distance to the best response (i.e., the child with the shortest distance (the winning child) receives rank 1, the child who is closest to winning receives rank 2, and so on). Then, we can compute the average ranks across rounds. Note that better choices lead to *lower* average ranks.

Gender and Age. Gender is indicated by a dummy that is equal to one if the child is female. Age is measured in dimensions of years, i.e., increasing age by one implies an increase in age of one year (still, age is measured referring to the exact date of birth of each child).

Fluid IQ. To proxy children's cognitive skills, we measured fluid IQ with 18 out of 36 available items (plus two example items to explain the test) from Raven's

[8] To double-check, one of the other experimenters wrote down the numbers for all children for all rounds during the game. In a very few cases (11 out of 1140), we used the number provided by the experimenter, mostly because of handwriting issues.

Colored Progressive Matrix test (Bulheller & Häcker, 2010). Each correctly solved item was weighted equally. The distribution of raw scores is shown in Figure A1 in the Appendix. To enable comparison of effect sizes, the variable was standardized to mean = 0 and SD = 1.

Understanding of the Game. After the instructions, each child had to explain the game back to the experimenter (see Sect. 2.3). The experimenter rated understanding for each of the five steps of the game on a 4-point scale, ranging from 4 = "Immediately completely and correctly explained" to 1 = "Not completely understood" (see Sect. F.2 in the Appendix). Understanding is calculated using the sum of the five steps, with a maximum value of 20.

We also collected some other measures that were not part of the present study. Finally, we checked whether parental consent for data use was available (which was the case for 114 out of 115 children), and analyzed the data using the statistical software Stata/SE 15.

3 Results and discussion

We first provide descriptive evidence to support that children successfully understood and played a strategic interaction game in the form of an experimental BCG. Then, we turn to the results regarding important determinants of performance in the BCG and discuss our findings. Finally, we report the results from our replication study with an adult sample.

3.1 Descriptive results

As noted, we had 114 participants in our sample, 63 of whom were female. Distributions over the classes and grades as well as summary statistics for fluid IQ and understanding of the game can be found in Table A1 in the Appendix.

To assess how well children understood the game, each child had to explain the five steps of the game to an experimenter. For each step, a child's level of understanding was rated by the experimenter with a maximum of four points, i.e., a child could achieve at most 20 points (see Sect. 2.3). Overall, 90 children (80%) were rated with 19 or 20 points, only five children (4%) received less than 17 points (see Table A1 in the Appendix).[9] Hence, experimenter ratings indicate a very high level of understanding of the game.

To analyze children's choices in the BCG, we compare the results from the present study with other data from experimental BCG studies. First, in Fig. 2 we plot mean choices (for median choices, see Tables 1 and A2 in the Appendix) over the 10 rounds and compare them with the choices from Nagel's seminal paper (1995, using sessions with $p = 2/3$ and sessions with $p = 1/2$) as well as with our own

[9] Excluding the children with ratings of understanding below 19 points does not affect our findings (see Sect. C in the Appendix).

replication with an adult sample (for details on the replication with adults, see Sect. 3.4).[10] Generally, we find that the average number chosen by children is decreasing over the 10 rounds, with the bulk of the decrease occurring in rounds 1–6. Thus, children's choices seem to converge toward the game-theoretic equilibrium over time, a finding that is well established in other studies with adult samples (e.g., Nagel, 1995; Duffy & Nagel, 1997; Ho et al., 1998). In addition, children's choices mimic both the level as well as the rate of decrease found in Nagel (1995)—however, the children sample is very close to the sessions with $p = 2/3$ in the Nagel study, indicating that children's choices start on a higher level (but develop with a similar rate of decrease). In other words, children seem to play a BCG with $p = 1/2$ using the new design proposed in our study in a very similar manner to the way that adults play the classical $p = 2/3$ experimental BCG.[11] Table 1 reports the detailed values comparing the first four rounds of the game across the four samples. Mean and median numbers for the children are largely comparable with the numbers from Nagel (1995) in the sessions using $p = 2/3$ (testing children's choices in rounds 1–4 against choices in the corresponding rounds in the Nagel $p = 2/3$ sample reveals no significant difference, Mann–Whitney U tests for each round, all $p > .164$).[12]

Second, to further support the notion that children played the new design of the experimental BCG in a way that is largely comparable with adults, we benchmark our data with results on the average "depth of reasoning" from Duffy and Nagel (1997). We apply Duffy and Nagel's definition of the depth of reasoning by calculating the depth of reasoning d for individual i in group j for round t solving:

$$number_{i,t} = median_{j,t-1} * p^{d_{i,t}}$$

Table 2 reports the values for our children sample as well as our replication with an adult sample and the comparable numbers from Duffy and Nagel (1997). Note that for this comparison, the game parameters of both designs match exactly, i.e., the data from Duffy and Nagel (1997) are also based on $p = 1/2$ and the median (subjects in their study were undergraduate university students). Table 2 clearly shows that, in all three samples, the majority of players chose numbers in the range of $d = 1$ (except for round 4 in the Duffy and Nagel sample), but in Duffy and Nagel slightly more mass lies on higher values of d compared with our design. Testing whether the distributions for levels of depth of reasoning are equal across samples reveals that children choose significantly different from the Duffy and Nagel sample in round 1 ($\chi^2 = 11.66, p = .040$) and in the subsequent rounds 2–4 (all $\chi^2 > 24.51$, all $p < .001$), with more probability mass in lower levels of depth of reasoning. Overall, adult university students in the Duffy and Nagel sample played with a slightly higher depth of reasoning than the children in our sample; however, considering the fact

[10] Note that for several reasons (see Sect. 2.3) we used the median, whereas Nagel (1995) used the mean to identify the winner in their experimental BCG.

[11] Also, the distribution of first-round choices is, by and large, in line with adult studies; see Figure A2 and, e.g., Nagel (1995) or Bosch-Domènech et al. (2002).

[12] Results for the adult sample are discussed in Sect. 3.4.

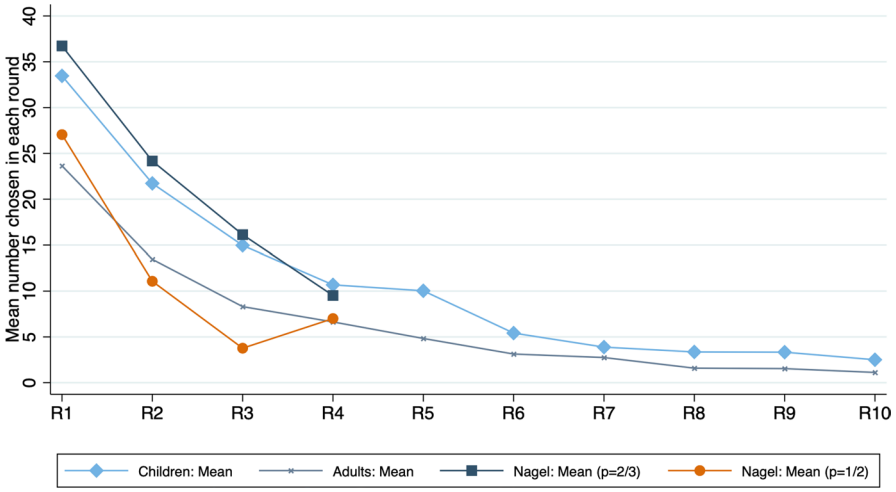

Fig. 2 Development of Mean Numbers Chosen over Rounds. *Notes:* This figure plots mean numbers chosen in each round for the present study (children and adult samples) compared with Nagel (1995), using averages of sessions with $p = 2/3$ and $p = 1/2$

Table 1 Comparison of Mean and Median Numbers with Nagel (1995)

	Children, $p = 1/2$		Adults, $p = 1/2$		Nagel, $p = 2/3$		Nagel, $p = 1/2$	
Round	Mean	Median	Mean	Median	Mean	Median	Mean	Median
1	33.5	28	23.6	21	36.7	33	27.1	17
2	21.7	20	13.5	11	24.2	18	11.1	7.25
3	15	11.5	8.3	7	16.1	12	3.8	3
4	10.7	8	6.6	4	9.5	8	7	0.755
RoD 1–4	0.66	0.69	0.72	0.75	0.74	0.76	0.76	0.95

Comparison of mean and median numbers chosen in our study (children and adult samples) with those in Nagel (1995); average values for sessions with $p = 2/3$ and $p = 1/2$, based on raw data provided by the author. RoD refers to the "Rate of Decrease" of the mean or median within a group over rounds 1–4. RoD $= (mean_{t=1} - mean_{t=4})/mean_{t=1}$, or the median, respectively. Reported rates of decrease are averages weighted by group size

that we played the BCG with children aged 9–11 years, the distributions of d are surprisingly comparable.[13]

[13] We also performed some additional analyses on determinants of game descriptives, both on the individual as well as on the group level (e.g., whether fluid IQ relates to choosing zero early in the game or other patterns of play). None of these tests yielded significant results.

Table 2 Comparison of depth of reasoning with Duffy and Nagel (1997)

Round	$d > 3$	$d = 3$	$d = 2$	$d = 1$	$d = 0$	$d < 0$
Children, $p = 1/2$, median						
1	.04	.02	.16	<u>.39</u>	.31	.08
2	.00	.01	.12	<u>.41</u>	.38	.08
3	.01	.03	.11	<u>.42</u>	.35	.08
4	.03	.02	.11	<u>.38</u>	.36	.11
Adults, $p = 1/2$, median						
1	.04	.06	.27	<u>.50</u>	.13	.01
2	.02	.03	.22	<u>.43</u>	.23	.08
3	.03	.04	.20	<u>.34</u>	.28	.11
4	.06	.06	.15	<u>.37</u>	.20	.17
Duffy and Nagel (1997), $p = 1/2$, median						
1	.07	.04	.30	<u>.43</u>	.11	.05
2	.02	.09	<u>.29</u>	<u>.29</u>	.16	.16
3	.02	.11	.32	<u>.38</u>	.09	.09
4	.07	.18	<u>.41</u>	.14	.05	.14

Comparison of the distribution of the average depth of reasoning in our study (children and adult sample) with the study by Duffy and Nagel (1997, p. 9). $d_{i,t}$ solves the equation of $number_{i,t} = median_{j,t-1} * p^{d_{i,t}}$ for individual i in group j for round t. We used 50 as the initial reference point. An individual i is categorized as, e.g., $d_{i,t} = 2$ if his or her $d_{i,t}$ lies within the interval with the boundaries $median_{j,t-1} * p^{2+1/2}$ and $median_{j,t-1} * p^{2-1/2}$. Modal categories are <u>underlined</u>

Combining the similarity of choices over the rounds with the fact that these choices were taken based on a correct understanding of the rules of the BCG, we derive our first main result:

Result 1 Using a new design of the experimental BCG, children aged 9–11 years are able to understand and play a strategic interaction game like the experimental BCG used in many other studies with adults. While children start with slightly higher numbers than adults, the rate of decrease and depth of reasoning are, by and large, comparable with values for adults.

3.2 Determinants of successful performance

Now, we turn to the question of whether fluid IQ, an important part of cognitive skills, is associated with high strategic interaction skills, i.e., successful performance in a BCG.[14] Importantly, all analyses presented are of a correlational nature; based

[14] We also collected data on a second set of abilities relevant for successful strategic interaction, namely perspective-taking abilities. We present some first, exploratory findings for these measures in Sect. D in the Appendix.

on our study design we cannot make any causal claim. Nevertheless, we believe that the analyses provide interesting insights into the importance of fluid IQ for choices in an experimental BCG, once the abstract version of the BCG is transformed into an easy-to-understand board game.

We present results from ordinary least squares (OLS) regressions; all models control for group fixed effects (i.e., we only look at differences within a group of five children playing the BCG) as well as gender and age. All standard errors are clustered at the group level. Gender is a binary variable; age is measured in years (but varies with a day-to-day precision between children). Fluid IQ is standardized to mean = 0 and SD = 1 to make interpretation easier (details on how we formed our outcome variables can be found in Sect. 2.4).

Fluid IQ was measured using a Raven's CPM test. We collected fluid IQ scores for all 114 children in our sample. Raw scores range from 5 to 18 and show large variation. The average score is 14.7, median score is 15, and the standard deviation is 2.8 (the full distribution of scores can be found in Figure A1 in the Appendix). Therefore, we believe that we have a meaningful measure of fluid IQ to proxy an important part of general cognitive skills.

In a first step, we analyze the link between fluid IQ and the probability of choosing a weakly dominated number, i.e., a number larger than 50.[15] We estimate the likelihood of choosing a weakly dominated number in the first round (in which no other behavior has been observed prior to the choice) and for all 10 rounds of the game, both using linear probability and probit models. As documented in Table A3 in the Appendix, there is a negative correlation between fluid IQ and the likelihood of choosing a weakly dominated strategy (statistically significant in three out of four specifications). The effect of fluid IQ is also large in size: moving from an average fluid IQ to a fluid IQ one SD above the average score substantially decreases the likelihood of choosing a weakly dominated strategy in the first round (from 27.1 to 16.7%, see column (2) in Table A3).[16]

Next, we investigate the relationship between the number chosen in the first round and fluid IQ. Previous studies show a negative relationship between cognitive skills and first-round choices in BCGs, i.e., individuals with higher cognitive skills choose closer to the game-theoretic equilibrium of zero (Burnham et al., 2009; Carpenter et al., 2013). However, in our setting there is no significant relationship between fluid IQ and and the number chosen in the first round (for more detailed findings and a short discussion of first-round choices, see Sect. B and Table A4 in the Appendix).

To analyze the predictive power of fluid IQ for strong strategic interaction skills, we first examine the number of coins a child has won, i.e., how many rounds the child won during the experimental BCG. In all subsequent analyses, we exclude results from the first round because here no prior interaction has taken place and the children cannot condition their choices on the observed behavior of their peers.

[15] We use a median BCG with $p = 1/2$. For all numbers larger than 50, a player would be equally or better off when choosing 50 instead; hence, numbers larger than 50 are weakly dominated.

[16] Note, however, that we can only use a subsample for this analysis because we control for group FE, and in eight out of the 23 groups no child chose a weakly dominated strategy in the first round.

Table 3, column (1), presents our findings for regressing the number of coins won on the dispositional characteristics of the children. It shows that neither gender, age, nor fluid IQ significantly explain variation. Hence, in this setting, cognitive skills, measured as fluid IQ, are not related to successful performance in the experimental BCG.

Using the number of coins (or rounds) won is easy and straightforward but there is a caveat: this measure disregards any difference in children's performance apart from being "the best" in a given round. For example, in these analyses, a child who fails to win a coin by only one step is considered equal to a child who misses half the median by 30 or more steps. Obviously, the latter child can be thought to have performed much worse than the first child. Thus, we also want to present an analysis accounting for the variations in performance between all five children in a group, not only an analysis of the winning child versus the non-winning children. To do this, we calculate the "distance to the best response", that is, how far a child is from the choice that would make him or her win the round, given the other children's choices (see Sect. 2.4 for details). Because this measure is very heterogeneous both across rounds as well as across groups, we *rank* children within groups and rounds based on their distance to the best response (i.e., as noted above in Sect. 2.4, the child with the shortest distance—the winning child—receives rank 1, the child with the second-shortest distance receives rank 2, and so on). We can then calculate an average rank for each child over rounds 2–10, with a "good performance" corresponding to a low average rank. We report the results from regressing average rank on the personal characteristics of the child in Table 3, column (2). Our findings confirm results from column (1): neither gender, nor age, nor fluid IQ are related to average rank over the rounds. In column (3), we analyze the average distance instead of the rank—results are very comparable.[17]

Next, we analyze children's behavior over the rounds to understand how children adapt their choices over time. To do so, we use the panel structure of our data setting the child as the panel variable and rounds 1–10 as the time variable. We cannot use conventional linear panel models in our setting because the unobserved panel-level effects will very likely be correlated with the lag of the dependent variable, in our case the number chosen in the previous round. We therefore use the Arellano–Bond estimator (Arellano & Bond, 1991) which basically instruments the first difference of interest (here choice in round t and $t-1$) with the second (and higher order) lag (e.g., choices in round 5 are instrumented by choices in rounds 3, 2, and 1).[18]

Table 4 presents results for the relationship between the number in round t and the following variables: numbers chosen in round $t-1$, winning number in round $t-1$, and position of the goblin in $t-1$. We see that (1) children on

[17] Estimating the models in Table 3 using school or class FE instead of group FE does not change the results.

[18] Note that in order to yield consistent estimates the Arellano–Bond estimator requires that the differenced unobserved time-invariant component is not correlated with the second (and higher order) lag of our outcome variable. We test this assumption in the full model (column (3) in Table 4) and find (weak) support: an Arellano–Bond test yields $p = .173$ for the test for autocorrelation of order 2; thus, we cannot reject the hypothesis that there is a zero second-order autocorrelation in our data set.

Table 3 Determinants of Successful Performance in the Game

	DV: Coins (R2–10) (1)	DV: Rank (R2–10) (2)	DV: Distance (R2–10) (3)
Female	0.072	0.033	−0.305
	(0.370)	(0.177)	(0.426)
Age in years	−0.200	0.120	−0.156
	(0.506)	(0.203)	(0.490)
Fluid IQ	0.251	−0.068	0.027
	(0.236)	(0.101)	(0.220)
Group FE	Yes	Yes	Yes
Observations	114	114	114

The results are based on OLS regressions. Standard errors in parentheses are clustered at the group level. *p< .10, **p< .05, ***p< .01

average choose lower numbers across rounds (the coefficient of number in round $t-1$ is < 1 in all models, confirming the results from our previous analyses), (2) children do—to some degree—"stick" to their number chosen in a previous round (the influence of choice in round $t-1$ is significant in all specifications), but (3) are more strongly influenced by the position of the goblin in the previous round (column (3) clearly shows that the position of the goblin in round $t-1$ has the strongest influence on the number chosen in round t). Because Arellano–Bond estimators do not allow for time-invariant controls, we conduct a median sample split in order to at least qualitatively analyze differences in learning behavior with respect to fluid IQ. Column (4) is based on the $n = 48$ children in our sample with a below-median score in the fluid IQ test, and column (5) is using the remaining $n = 66$ children (at or above median IQ score). Results from these two models point to two potential mechanisms how fluid IQ might affect learning behavior: First, it seems that children with lower fluid IQ tend to more strongly "stick" to their choices from the previous rounds (comparing the coefficients of Number Chosen in $t-1$). Second, the findings seem to suggest that both groups pay attention to the position of the goblin in round $t-1$ when choosing a number in round t, but that only children with a higher fluid IQ do potentially account for the position of the winning child (although this effect is statistically not significant). The latter strategy, however, seems important for successful game performance because when forming beliefs about the other players' next round choices it is potentially important whether the winning child was above or below the goblin's position (i.e., 50% of the median number). Of course, given sample size and power, these findings have to be interpreted cautiously and can only point to interesting mechanisms for further research.

Finally, we also estimate how a child's depth of reasoning is related to individual characteristics. Note that a high level of depth of reasoning in itself is not necessarily a predictor of good performance in the game—children could display an excessively high depth of reasoning and "outsmart themselves" by choosing numbers that are too low (cf. Kocher & Sutter, 2005). Table A5 in the Appendix reports that there

Table 4 Prior choices and learning over rounds

	DV: Number Chosen in Round t				
	(1)	(2)	(3)	(4)	(5)
				Below-median IQ	Above-median IQ
Number chosen in $t-1$	0.435***	0.231***	0.189***	0.277***	0.147**
	(0.036)	(0.060)	(0.057)	(0.078)	(0.060)
Winning number in $t-1$		0.623***	0.167	−0.094	0.295
		(0.141)	(0.280)	(0.504)	(0.330)
Goblin position in $t-1$			0.662*	0.784	0.560
			(0.358)	(0.650)	(0.406)
Observations	912	912	912	384	528

The results are based on a linear dynamic panel models using the Arellano–Bond estimator (Arellano & Bond, 1991). Models (1)–(3) are based on the full sample of $n = 114$ children. Models (4) and (5) are based on a median split in fluid IQ scores, with $n = 48$ children in the below-median fluid IQ group and $n = 66$ in the at- or above-median fluid IQ group. Standard errors in parentheses are clustered at the individual level

*$p < .10$, **$p < .05$, *** $p < .01$

is no relationship of gender, age, or fluid IQ with a child's average depth of reasoning across rounds. Thus, we can conclude our second main result:

Result 2 In our new design of the experimental BCG, cognitive skills—measured as fluid IQ—are neither related to first round choices nor to successful performance but only predict the choice of weakly dominated strategies. Similarly, fluid IQ is not associated with a child's average depth of reasoning.

Taken together, our findings indicate that strategic interaction skills in the new design of the experimental BCG are not linked to gender or age (within our age range of 9–11 years). Fluid IQ is only relevant in predicting whether children choose weakly dominated strategies but is not associated with first round choices, successful performance, or higher depth of reasoning in the game. To support the stability of our findings, we conducted several robustness checks (excluding children with low understanding, excluding weakly dominated choices, estimations without group-fixed effects). In all versions, our findings remain stable (for details, see Section C in the Appendix).

3.3 Discussion

Previous studies have generally demonstrated positive links between cognitive skills and lower entries in the experimental BCG (Burnham et al., 2009; Brañas-Garza et al., 2012; Carpenter et al., 2013). We also find a link between cognitive skills, measured as fluid IQ, and choosing weakly dominated strategies, replicating findings from Burnham et al. (2009, p. 172). Yet, Burnham et al. (2009) also document a

relationship between higher cognitive skills and lower numbers chosen in a one-shot BCG which we cannot replicate (first-round choices are not linked to fluid IQ, see Table A4 in the Appendix). We believe that the specific measure for cognitive skills could be an aspect that helps explain this: Brañas-Garza et al. (2012) use a Raven's IQ test, as we do in our study, and also find no relationship between fluid IQ and choices in an experimental BCG (yet, they do report a significant link to the Cognitive Reflection Task). Moreover, our sample consists of children aged 9–11 years and it is possible that, for this age group, other abilities simply matter more than IQ; however, this explanation is made less likely by the fact that in our replication using an adult sample, there is no significant relationship between fluid IQ and successful performance (see Sect. 3.4). This leads us to the explanation that we consider most plausible. The difference between our results and previous findings could be driven by the fact that in all these studies, the instructions for the BCG were abstract and, therefore, cognitive skills were more important (or even a prerequisite) for understanding the mechanisms of the game. In our setting, the instructions are far more concrete and the game itself has a visual and spatial representation. In other words, choices and their consequences are mapped into concrete and observable operations. Thus, our speculative hypothesis is that the new design of the experimental BCG lowers the importance of cognitive skills for successful performance in the game. Note that by removing the requirement to translate abstract instructions into concrete operations, we can study actual behavior in strategic interaction settings in a much more focused way. Indeed, real-world strategic interaction is often characterized by repetition, observable behavior, and concrete outcomes, as well as possibilities to learn from one's choices. Hence, removing (or lowering the demand for) this abstract component from the experimental BCG might actually increase external validity for real-world strategic interaction.

There are two methodological challenges with the lack of significant relationships between fluid IQ and successful performance. First, this could be due to restricted or limited variance. If, for example, classes or groups were very homogeneous with respect to fluid IQ levels, this could (partially) explain why there is no significant link between fluid IQ and performance. To analyze this concern, we checked the variance of the results from the Raven's Matrices task within our sample. For the whole sample, the variance in raw scores for fluid IQ amounts to 8.0 points. Calculating the within-class variance and then averaging over these within-class variances for all classes (weighted by class size) results in an average variance at the class level of 6.4 points.[19] Because we assigned children randomly to groups (within a class), we expect that the average variance at the group level would not differ from that at the class level. Indeed, the average within-group variance amounts to 6.6 points. Finally, we can compare the variance in our study with figures from a different study using the same test for fluid IQ (Berger et al., 2020). In four different testing waves with a sample of more than 500 German primary schoolchildren aged 7–9

[19] Fluid IQ scores are positively correlated with age. Given that we cover an age range of 9–11 years and that classes are homogeneous in terms of age, it is not surprising that the variance for the whole sample is somewhat larger than the average variance at the class level.

years, this study finds variances of 8.3, 7.1, 8.1, and 6.3 points, respectively. Thus, the variance of fluid IQ in our sample (and within our groups) is substantial and in line with other, much larger samples of schoolchildren. An alternative way of testing for within-group-restricted variance as a potential explanatory factor is to estimate the OLS models from Table 3 without group-fixed effects. In doing so, we exploit the full distribution of IQ scores within our sample (but also lose control over other factors varying between groups). We report these estimations in Table A9 in the Appendix; there is no significant link between fluid IQ and any of our measures of successful performance when excluding group-fixed effects.

Second, the lack of a significant relationship between fluid IQ and successful strategic interaction could be due to limited statistical power of our study. However, comparing our results for the first round choices with findings from Burnham et al. (2009, they use a BCG with $p = 1/2$ and choices between 0 and 100), we see that they report an effect of −9.67 in first-round choices for a one standard deviation increase in cognitive skills (p. 173). In contrast, if we regress the number chosen in round 1 on gender, age and fluid IQ (clustering standard errors at the group level), the 95% confidence interval for a one standard deviation increase in fluid IQ on the number chosen in round 1 ranges from −7.26 to 4.29; this suggests that we can basically rule out effects of fluid IQ in the size found by Burnham et al. (2009). In addition, our exploratory analysis on the link between perspective-taking abilities and successful performance in the BCG (see Sect. D in the Appendix) indicates that for other individual characteristics (even a binary one), our study seems to be sufficiently powered to identify statistically significant relationships.

Finally, we do not identify any significant relationship between age and successful strategic interaction. In principle, an increase in strategic interaction skills with age could be anticipated (e.g., Brosig-Koch et al. (2015) show that children's ability to reason backward clearly improves from the age of 6 years onward, and Charness et al. (2019) show with a sample of children between 3 and 11 years old that Theory of Mind increases considerably with age). On the other hand, Czermak et al. (2016) find no substantial effects of age in strategy games and conclude that their results suggest that "strategic decision-making is fairly well developed at an age of 10 years and hardly changes in subsequent years" (p. 270). Potentially, this conclusion could already apply at the age of 9 years, which is the lower bound of age in our sample of children. However, our results regarding age must be interpreted with caution because (1) we study a rather small age range of only two years, (2) we only compare children within groups who are even more homogeneous in age than the full sample (because groups were randomly drawn from the same class), and (3) our distribution of age is not linear because we study a cohort of third and fifth graders (i.e., there are no fourth graders in the sample). Hence, identifying age effects in such a setting is challenging and the absence of significant age differences in successful strategic interaction in our study should not be interpreted as evidence of the absence of development in strategic interaction skills within this age range.

3.4 Replication with an adult sample

When we tested the new design of the experimental BCG with our sample of children and compared it with the previous findings in the literature, we changed two factors simultaneously: the design of the game and the sample of participants. To provide the "missing piece", we replicated our study using the new design of the experimental BCG but with an adult sample of university students.

Experimental Design. We recruited $n = 120$ participants, 60% of whom were female, with a mean age of 22.6 years (see Table A12 in the Appendix for details). The experiment was conducted in the MABELLA (Mainz Behavioral and Experimental Laboratory). The sessions were combined with another experiment but mirrored the basic structure of the study with children: first, all participants within a session ($n = 10$) were seated at separate tables in a large room and filled out questionnaires and tests (including a (short) version of Raven's Matrices for adults). Subsequently, two groups of five adults (randomly assigned) each went to a separate room with an experimenter to play the new design of the experimental BCG. The only difference with the study with children was that adults did not receive one-to-one instructions but were instructed as a group (also, they did not have to explain the game back to the experimenter). After the BCG, participants were paid anonymously in a separate room. We conducted 12 sessions with two groups each, sessions lasted for 70–90 min in total. Average payoff was EUR 15.45, including a show-up fee of EUR 5. The experimental BCG was incentivized, with the winner of a randomly drawn round receiving EUR 20. The fluid IQ test was not incentivized.

Results. Figure 2 and Table 1 show that choices by adults start at a lower level than those by children and remain consistently lower in subsequent rounds (Mann–Whitney U test for rounds 1–4 for each round, all $p < .0001$). More detailed information on adults' choices can be found in Table A13 and Figure A3. Benchmarking adults' choices with choices in the study by Nagel (1995) suggests that our adult sample playing the BCG in the new design behaves very similarly to the adult sample in Nagel's study playing the classical BCG with $p = 1/2$. Comparing the numbers chosen in the first round for our adult sample and for the Nagel $p = 1/2$ sample reveals no significant difference (Mann–Whitney U test, $p = .899$). However, in rounds 2–4 the Mann–Whitney U tests are significant, suggesting that participants in the Nagel $p = 1/2$ sample choose lower numbers than those in our adults sample (Mann–Whitney U tests for each round, all $p < .032$), which is in line with the notion that median choices decrease somewhat faster in the classical BCG, see Table 1.[20]

When comparing the distributions of depth of reasoning in rounds 1–4 in Table 2, the majority of adults display a $d = 1$. If we compare the distributions of levels of d, we see that our adult sample is significantly different from the children sample in round 1 ($\chi^2 = 23.10, p < .001$), with adults having more probability mass in higher levels of reasoning, as one would expect. In later rounds, however, the picture is

[20] A potential difference driving this could be the fact that Nagel (1995) used the mean, whereas we used the median in our study to identify the winner.

mixed.[21] Comparing the depth of reasoning in our adult sample with the adult sample in Duffy and Nagel (1997), we find that for round 1 the distributions of d are not significantly different ($\chi^2 = 4.95, p = .422$). In subsequent rounds, adults in our sample show lower levels of depth of reasoning than the adults in Duffy and Nagel (for rounds 2–4, all $\chi^2 > 9.82$, all $p < .080$). Overall, levels of d in our adult sample are slightly higher than in our children sample and slightly lower than in the adult sample by Duffy and Nagel (1997), which places the distribution of our adult sample playing the new design of the BCG *between* the children sample playing the new design of the BCG and the adult sample by Duffy and Nagel (1997) playing the classical design.

In total, the replication of the new design of the experimental BCG with an adult sample shows that this design can be used successfully to study strategic interaction with adults:

Result 3 Adults playing the new design of the experimental BCG behave in a way that is largely comparable with adult behavior in the classical BCG. The average numbers chosen and the rate of decrease are very similar to those chosen by other adult samples. The average depth of reasoning is higher than that in our children sample and slightly lower than that in the classical BCG.

Although this was not the focus of our replication study, we also conducted a parallel analysis of the link between cognitive skills, measured as fluid IQ, and successful performance for adults. Results can be found in Table A14. Similar to the children sample, successful performance is not related to fluid IQ. This indicates that the new design might indeed place lower demands on cognitive skills by making the game less abstract and easier to understand (see Sect. 3.3). In addition, we document a substantial gender difference in the adult sample. Women perform significantly worse than men when looking at the number of coins won and the average rank.[22] Taken together, the replication study using an adult student sample generally confirms that the new design of the experimental BCG can also be used with adults. Further investigating determinants of successful performance and gender differences in strategic interactions for adults as well as skills that have a *causal* effect on successful performance appear to be promising avenues for further research.

[21] The distributions of d are (marginally) significantly different in round 2 ($\chi^2 = 9.87, p < .079$), not significantly different in round 3 ($\chi^2 = 6.38, p < .271$), and significantly different in round 4 ($\chi^2 = 11.37, p < .044$).

[22] When analyzing how the adult sample adapts choices over time (compared to Table 4), we find qualitatively very similar results. This similarity is largely in line with results of testing for differences in the rates of decrease from Table 1: Testing for equality in central tendencies between the rates of decrease for children vs. adults reveals no significant difference for the decrease in mean numbers (Mann–Whitney U test, $p = .349$) but a significant difference for the decrease in median numbers (Mann–Whitney U test, $p = .045$).

4 Conclusion

This paper introduces a new design of the experimental BCG. We use this new design for the first-ever study conducting a BCG with groups of children and demonstrate that children are capable of understanding and playing this BCG. This allows for a wide range of applications in studying the skill formation process for strategic interaction. Moreover, our findings demonstrate that an important part of cognitive skills, namely fluid IQ, is not significantly related to successful performance in this strategic interaction setting, opening up the question which skills are important to succeed in strategic interaction. At the same time, our new design allows for research designs focusing on the development of strategic interaction skills (and, potentially, its determinants) starting already at young age. Finally, in the implementation of the new BCG design with adults we find results largely in line with behavior in the classical BCG, suggesting that our new BCG design can also be used with adult samples.

In future research, it would be promising to extend the age range studied, for example, to children entering school or adolescents. This could improve our understanding of how the ability to strategically interact in groups develops with age. Moreover, in order to advance our understanding of important determinants of successful strategic interaction it seems promising to include measures of abilities in the area of perspective-taking and empathy (see Sect. D in the Appendix for some first, exploratory evidence). A different perspective could use longitudinal data to analyze predictors of successful strategic interaction, i.e., which skills and abilities are the building blocks of this complex ability? Relatedly, which background characteristics are linked to strategic interaction skills? For example, the detection of an early gap in strategic interaction skills based on socioeconomic background (controlling for cognitive skills and other important abilities) would contribute to our understanding of the intergenerational transmission of strategic interaction skills as well as the origins of socioeconomic inequalities. Finally, causal evidence, e.g., using priming and/or cognitive load paradigms, or even targeted interventions could provide further insights into the importance of various skills as building blocks of strategic interaction skills.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10683-021-09713-y.

Acknowledgements We are grateful to the editor and two anonymous referees for very helpful feedback. In addition, we would like to thank our colleague Aleksa Kaurin for teaming up in the data collection as well as Helen Chairopoulou, Kristina Jakob, Helena Rohlik, Marion Schmale, and Vivien Voigt for excellent research assistance. We also thank Eirik Berger, Matthias Heinz, Fanny Landaud, Matthias Sutter, Bertil Tungodden, Justin Valasek, and Constantin Weiser for helpful comments as well as Rosemarie Nagel for providing raw data from her experiment. In addition, the paper benefited from feedback at the ESA Conference 2019 in Abu Dhabi, the Workshop on Microeconomics in Lueneburg, the EEG seminar at MPI Bonn, and NCBEE 2019 in Kiel. This project was generously supported in part by the Jacobs Foundation and the research unit Interdisciplinary Public Policy at the University of Mainz (IPP Mainz). Henning Hermes gratefully acknowledges financial support from the German National Academic Foundation (Studienstiftung des deutschen Volkes) and the Research Council of Norway through its Centers of Excellence Scheme, FAIR (Projects Nos. 262675, 262636, and 250170F10).

Funding Open access funding provided by Norwegian School Of Economics.

References

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies, 58*(2), 277–297.

Berger, E. M., Fehr, E., Hermes, H., Schunk, D., & Winkel, K. (2020). The impact of working memory training on children's cognitive and noncognitive skills. IZA Discussion Paper No. 13338. Bonn: IZA.

Birch, S., Li, V., Haddock, T., Ghrear, S., Brosseau-Liard, P., Baimel, A., & Whyte, M. (2017). Chapter six—Perspectives on perspective taking: How children think about the minds of others. In B. Janette (Ed.), *Advances in Child Development and Behavior* (Vol. 52, pp. 185–226). Benson: JAI.

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

Bosch-Rosa, C., & Meissner, T. (2020). The one player guessing game: a diagnosis on the relationship between equilibrium play, beliefs, and best responses. *Experimental Economics*, *23*, 1129–1147. https://doi.org/10.1007/s10683-020-09642-2.

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

Brandts, J., & Cooper, D. J. (2006). A change would do you good .... an experimental study on how to overcome coordination failure in organizations. *American Economic Review, 96*(3), 669–693.

Braun, S., Dwenger, N., Kübler, D., & Westkamp, A. (2014). Implementing quotas in university admissions: An experimental analysis. *Games and Economic Behavior, 85,* 232–251.

Brocas, I., & Carrillo, J. D. (2018a). Iterative dominance in young children: Experimental evidence in simple two-person games. *Journal of Economic Behavior & Organization*, *179*, 623–637. https://doi.org/10.1016/j.jebo.2018.06.006.

Brocas, I., & Carrillo, J. D. (2018b). The determinants of strategic thinking in preschool children. *PLoS ONE, 13*(5), 1–14.

Brocas, I., & Carrillo, J. D. (2019). *Steps of reasoning in children and adolescents*. Technical report.

Brocas, I., & Carrillo, J. D. (2020). The evolution of choice and learning in the two-person beauty contest game from kindergarten to adulthood. *Games and Economic Behavior, 120,* 132–143.

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

Bulheller, S., & Häcker, H. O. (2010). *Coloured Progressive Matrices (CPM). Deutsche Bearbeitung und Normierung nach J. C. Raven*. Frankfurt: Pearson Assessment.

Burnham, T. C., Cesarini, D., Johannesson, M., Lichtenstein, P., & 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.

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

Castillo, M., Jordan, J. L., & Petrie, R. (2018). Children's rationality, risk attitudes and field behavior. *European Economic Review, 102,* 62–81.

Charness, G., List, J. A., Rustichini, A., Samek, A., & De Ven, J. V. (2019). Theory of mind among disadvantaged children: Evidence from a field experiment. *Journal of Economic Behavior & Organization, 166,* 174–194.
Cooper, D. J., & Kagel, J. H. (2005). Are two heads better than one? Team versus individual play in signaling games. *American Economic Review, 95*(3), 477–509.
Czermak, S., Feri, F., Glätzle-Rützler, D., & Sutter, M. (2016). How strategic are children and adolescents? Experimental evidence from normal-form games. *Journal of Economic Behavior & Organization, 128,* 265–285.
Duffy, J., & Nagel, R. (1997). On the robustness of behaviour in experimental 'beauty contest' games. *The Economic Journal, 107*(445), 1684–1700.
Egan, D., Merkle, C., & Weber, M. (2014). Second-order beliefs and the individual investor. *Journal of Economic Behavior & Organization, Empirical Behavioral Finance, 107*(Part B), 652–666.
Fe, E., Gill, D., & Prowse, V. (2019). *Cognitive skills, strategic sophistication, and life outcomes*. Technical report. Competitive Advantage in the Global Economy (CAGE).
Grosskopf, B., & Nagel, R. (2008). The two-person beauty contest. *Games and Economic Behavior, 62*(1), 93–99.
Hanaki, N., Koriyama, Y., Sutan, A., & Willinger, M. (2019). The strategic environment effect in beauty contest games. *Games and Economic Behavior, 113,* 587–610.
Harbaugh, W. T., & Krause, K. (2000). Children's altruism in public good and dictator experiments. *Economic Inquiry, 38*(1), 95–109.
Harbaugh, W. T., Krause, K., Liday, S. G., & Vesterlund, L. (2003a). Trust in children. In E. Ostrom & J. Walker (Eds.), *Trust and reciprocity: Interdisciplinary lessons for experimental research* (pp. 302–322). New York: Russell Sage Foundation. https://www.jstor.org/stable/10.7758/9781610444347.
Harbaugh, W. T., Krause, K., & Liday, S. J. (2003b). *Bargaining by children*. SSRN Scholarly Paper ID 436504. Rochester, NY: Social Science Research Network.
Ho, T.-H., Camerer, C., & Weigelt, K. (1998). Iterated dominance and iterated best response in experimental "p-beauty contests''. *The American Economic Review, 88*(4), 947–969.
Kautz, T., Heckman, J., Diris, Ron, W., Bas, T., & Borghans, L. (2014). *Fostering and measuring skills*. OECD education working papers. Paris: Organisation for Economic Co-operation and Development.
Kocher, M. G., & Sutter, M. (2005). The decision maker matters: Individual versus group behaviour in experimental beauty-contest games. *The Economic Journal, 115*(500), 200–223.
Kocher, M. G., & Sutter, M. (2006). Time is money-time pressure, incentives, and the quality of decision-making. *Journal of Economic Behavior & Organization, 61*(3), 375–392.
Murnighan, J. K., & Saxon, M. S. (1998). Ultimatum bargaining by children and adults. *Journal of Economic Psychology, 19*(4), 415–445.
Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review, 85*(5), 1313–1326.
Piaget, J. (1962). The stages of the intellectual development of the child. *Bulletin of the Menninger Clinic, 26*(3), 120–128.
Rangvid, J., Schmeling, M., & Schrimpf, A. (2013). What do professional forecasters' stock market expectations tell us about herding, information extraction and beauty contests? *Journal of Empirical Finance, 20,* 109–129.
Sher, I., Koenig, M., & Rustichini, A. (2014). Children's strategic theory of mind. *Proceedings of the National Academy of Sciences, 111*(37), 13307–13312.
Sutter, M. (2005). Are four heads better than two? An experimental beauty-contest game with teams of different size. *Economics Letters, 88*(1), 41–46.
Sutter, M., Zoller, C., & Glätzle-Rützler, D. (2019). Economic behavior of children and adolescents: A first survey of experimental economics results. *European Economic Review, 111,* 98–121.
Weber, R. A. (2003). 'Learning' with no feedback in a competitive guessing game. *Games and Economic Behavior, 44*(1), 134–144.
Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655–684.