Exploring the capability to reason backwards: An experimental study with children, adolescents, and young adults

Jeannette Brosig-Koch, Timo Heinrich, and Christoph Helbach*

This is the first study investigating the development of the capability to reason backwards in children, adolescents, and young adults aged 6 to 23 under controlled laboratory conditions. The experimental design employs a modified version of the race game. As in the original game, subjects need to apply backward analysis in order to solve the games. We find that subjects' capability to reason backwards improves with age, but that this process systematically differs across genders. Our repetition of the games indicates that differences exist also in learning between age groups and across genders.

Keywords: backward analysis, learning, age effects, experimental economics, children

JEL-No.: C72, J13, C91

* Faculty of Economics and Business Administration, University of Duisburg-Essen, Universitätsstrasse 12, 45117 Essen, Germany; e-mail: jeannette.brosig@uni-due.de (corresponding author), timo.heinrich@ibes.uni-due.de, christoph.helbach@ibes.uni-due.de.

1 Introduction

Dynamic decisions are of importance in many areas of daily life (e.g., sequential negotiations, health prevention, making arrangements for retirement, or investing in education). As long as one can assume that there is a last period, people need to apply some form of backward reasoning in order to calculate their optimal decision. At the least, backward induction is a fundamental assumption in modeling such decisions in economics. But are people capable of reasoning backwards?

Several experimental studies have reported what appear to be failures to apply backward induction. For example, in centipede games very few subjects play the subgame-perfect equilibrium strategy suggested by game theory and end the game at the first node (see McKelvey and Palfrey, 1992, Fey et al., 1996, Nagel and Tang, 1998, Parco et al., 2002, Rapoport et al., 2003, Bornstein et al., 2004). As this solution depends on common knowledge of selfishness and rationality, explanations such as the existence of social preferences or limited knowledge of rationality have been proposed. In a recent field experiment with chess players, Levitt et al. (2011) contrast the behavior in the centipede game with that in the race game. In the latter game the equilibrium can also be found by backward analysis,[1] but its game-theoretic solution is more robust than the solution of the centipede game. In the race game two players alternate in choosing numbers between 1 and an integer *k*. All chosen numbers are added up and the player who chooses a number that makes the sum equal to an integer *m* wins. Using this game has the advantage that the optimal strategy does not depend on beliefs about other players and, since it is a constant sum, winner-take-all game, it also does not depend on distributional or efficiency concerns. Some chess players in the study by Levitt et al. (2011) prove to be quite sophisticated in solving the race game with *k* equal to 9 or 10 and *m* equal to 100. But Levitt et al. (2011) observe no systematic relationship to the behavior in the centipede game. They conclude that the rather late stops in the centipede game are not driven by the subjects' inability to reason backwards.

In light of this evidence, two basic observations can be made: (i) a non-negligible share of people appears to be able to reason backwards; (ii) in the centipede game the frequency of equilibrium play depends on information about the opponents.

[1] As Dufwenberg et al. (2010) and Levitt et al. (2011) point out, the race game does not require backward induction in a *strict* sense as a player does not need to solve for her opponent's optimal choice. To acknowledge the different approach that is required to solve the race game, we follow Gneezy et al. (2010) and use the terms "backward analysis" or "backward reasoning".

In this study we build on observation (i) and ask how differences in the ability to apply backward analysis evolve. We extend the previous research by focusing on the development of this capability among different age groups. In particular, using modified versions of the race game, we compare how these games are solved by children, adolescents, and young adults aged 6 to 23. Additionally, we study whether there are differences between these age groups regarding the improvement of their performance. Exploring the development of backward analysis skills is interesting in its own right. However, insights on these abilities may also have consequences for modeling inter-temporal decisions, which are often considered a game against a future self (see, e.g., Laibson, 1997, and Diamond and Köszegi, 2003). Because many fundamental inter-temporal decisions are made early in life, for example planning a family or investing in education, it is important to learn how young people make such decisions.

Observation (ii) underlines the importance of selecting an appropriate experimental design to isolate the factors that influence the application of backward analysis. Because our focus is on the ability to reason backwards, we follow Levitt et al. (2011) in choosing the race game as an experimental paradigm. Additionally, to increase comparability between age groups, we opt for a design in which all subjects face the same computerized opponent (that plays the equilibrium strategy whenever possible).

The paper proceeds as follows. The next section surveys related literature from which we derive the hypotheses tested in the experiment. Section 3 presents the race games that were used in our study and section 4 describes the experimental design. The results are provided in section 5. Section 6 concludes.

2 Related literature and hypotheses

Our analyses are based on the race game which has been introduced to behavioral research in studies by Burks et al. (2009), Dufwenberg et al. (2010), and Gneezy et al. (2010). Employing two race games with k equal to 3 and 4 and m equal to 14 and 16, respectively, Gneezy et al. (2010) study whether, and how fast, subjects learn to reason backwards.[2] They observe that subjects only seem to apply this method after experiencing defeat. Dufwenberg et al. (2010) focus on learning transfers across games with k equal to 2 and m equal to 6 and 21. They report that experience with the shorter game improves performance in the longer one, though subjects seem to work "this analytic solution out in steps" (p. 141). Similar to the findings by

[2] In their notation the games are called $G(15, 3)$ and $G(17, 4)$.

Gneezy et al., their results suggest that cognitive limitations prevent subjects from reasoning backwards right from the beginning. In fact, based on a sample of 1,000 trainee truckers, Burks et al. (2009) find a significantly positive relationship between performance in a race game (referred to as Hit 15), IQ measured by a nonverbal IQ test (Raven's matrices), and a test of quantitative literacy. Similarly, conducting an online survey with 422 students, Carpenter et al. (2013) observe a strong correlation between performance related to Hit 15 and two common measures for cognitive ability (i.e., Frederick's CRT and college entrance exam scores from the Scholastic Achievement Test and the American College Test). The results by Burks et al. also reveal that the ability to solve the race game is positively related to patience, to the willingness to take calculated risks, and to the truckers' perseverance on the job, among other things. This relationship to other economically important behavioral traits and behavior in the field further emphasizes the value of finding out more about the capability to reason backwards.

Based on previous research on the race game, we formulate two basic hypotheses regarding the capability to reason backwards which we test in our experiment. First, we expect subjects to solve initially at least some of the race games by reasoning backwards:

Hypothesis 1: Subjects are able to reason backwards, but to a limited degree.

Second, because previous studies suggest that subjects learn to reason backwards when games are played repeatedly, we also formulate a second hypothesis:

Hypothesis 2: The ability to reason backwards improves with repetition.

As there is no evidence on the occurrence of this capability in children and adolescents, we expect the two hypotheses to hold for all age groups employed in our study. Though, there might be differences across these groups regarding average performance in the race game and its improvement.

In recent years, the influence of age on decision-making has started to gain attention in economics. To our knowledge, this is the first paper that studies backward analysis in children and across age groups. Most closely related to our study is the experimental research on the development of rationality (which is an important assumption in many models of strategic decision-making) and strategic behavior with age in an economic context. Harbaugh et al. (2001) test whether children exhibit rational choice behavior. They find that choices of 6th graders (about 11 years old on average) are as consistent as choices of undergraduates (about 21 years), while 2nd graders (about 7 years) decide inconsistently more often. Czermak et al. (2011) investigate the development of strategic behavior. However, they consider a smaller

age range of 10 to 17 year olds as well as static two-person games. They observe no influence of age on the likelihood of behaving strategically. The results of both studies suggest that rational behavior develops early, but does not change much thereafter.

Related developmental research points to an age-related increase of cognitive abilities. For example, Schneider et al. (2008) present the results of non-verbal IQ tests applied to 10, 12, 18, and 23 year old subjects in the Munich Longitudinal study of the Ontogenesis of Individual Competencies (LOGIC study).[3] They find that the tested abilities increase with age and that this increase becomes smaller at the latter measurement points. Stern (1999) focuses on mathematical competencies tested in the LOGIC study.[4] She reports that "for all mathematical tests presented at different age levels, an increase in performance level was obtained" (p. 161). Moreover, based on an updated data set, Stern (2008) finds significant gender differences in favor of males at all age levels; the effect size varies among age groups, however. The results of a recent meta-analysis on mathematics performance conducted by Lindberg et al. (2010) also point to age-related gender effects. The analysis is based on data from 242 studies published between 1990 and 2007 and reveals that differences between males and females "were negligible in elementary-school and middle-school-aged children and reached a peak […] in high school. The gender difference then declined for college-age samples and adults." (Lindberg et al., 2010, p. 1128). In addition to their meta-analysis, Lindberg et al. (2010) investigate large data sets on math performance based on probability sampling of U.S. adolescents (i.e., the National Longitudinal Survey of Youth, The National Educational Longitudinal Study, the Longitudinal Study of American Youth, and the National Assessment of Educational Progress). Consistent with their results from the meta-analysis, the authors observe that gender differences (favoring males) are somewhat higher in high school than in elementary or middle school. Lindberg et al. (2010) further report that observed gender differences seem to vary with the problem type (presence of multiple choice, short answer, and open ended questions) and the mathematical contents (proportion of algebra items and measurement items), however. Moreover, in a related study based on data from state assessment (including California, Connecticut, Indiana, Kentucky, Minnesota, Missouri,

[3] The nonverbal IQ tests include Cattell's Culture Free Intelligence Test and Arlin's Test of Formal Reasoning. The first test assesses fluid intelligence and requires subjects to identify and complete series of geometric figures, to classify and differentiate geometric figures, to complete matrix figures, and to identify proportions and relations of geometric areas. The second test focuses on operational reasoning, i.e. on the transition between concrete and formal operations.

[4] These competencies involve word problems dealing with the comparison of sets, numerical problems that require using elaborated strategies, and problems dealing with proportions (see Stern, 1999).

New Jersey, New Mexico, West Virginia, and Wyoming), Hyde et al. (2008) observe no gender difference in performance at any grade level through grade 11.

Some support for age-related gender effects has also been provided by developmental research testing individual planning abilities (which are related to different forms of cognitive flexibility; see McCormack and Atance, 2011). Among others, studies on the Tower of Hanoi or Tower of London tasks suggest that planning abilities gradually increase with age and reach a maximum at young adult age (see, e.g., Bishop et al., 2001, for the Tower of Hanoi task, and Unterrainer et al., 2014, for the tower of London task). In these tasks, subjects have to rearrange disks of different sizes or balls of different colors from one given state to a goal state. Subjects' challenge is to reach this goal state with the minimum number of moves while adhering to some predefined rules of rearrangement. Based on the Tower of Hanoi task, Bishop et al. (2001) provide a comparison of male and female performance for each of the tested age groups separately. They find gender differences favoring males only for 9 to 10 year olds and for 11 to 12 year olds, but not for 7 to 8 year olds, 13 to 15 year olds, and adults. Unterrainer et al. (2014) do not report on gender effects in their test of behavior in the tower of London task.

The findings from related economic experiments and developmental research suggest that the ability to reason backwards improves with age. Developmental research on mathematical performance and on the tower of Hanoi task further indicates that there are age-related gender effects. Gender effects are not reported in all studies and do not uniformly refer to the same age groups, however. Accordingly, we formulate a broader third hypothesis on the effect of age:

> *Hypothesis 3*: The ability to reason backwards improves with age, but the improvement differs across genders.

3 Games

To study the capability to reason backwards across age and gender, we employ six different race games $G(m, k)$ in which two players alternate in choosing numbers between 1 and (k =) 4. The player who chooses a number that makes the sum of all chosen numbers equal m wins. The six games only vary regarding m, which takes the values 19, 3, 29, 8, 11, and 21,

respectively. In order to identify improvement in performance (i.e. in order to test Hypothesis 2), the six games are played twice in identical order.[5]

The race game can be solved by backward analysis. In order to reach a sum equal to m in her last move, player 1 needs to reach m-(k+1) on her second to last move. This way player 2 has no chance of reaching m on his last move. To be able to reach m-(k+1) on her second to last move, player 1 needs to secure position m-2(k+1) on the move before. More generally, she has to secure position m-(n-1)(k+1) in position n where n refers to the number of her moves that remain. Accordingly, the first mover can win all race games except those where m is divisible by (k+1). This implies that our games require 0 to 5 steps of backward analysis to be solved and that all of them can be won by the first mover.

4 Experimental design

The experiment was conducted with subjects of six different age groups. Subjects were recruited from an elementary school (with about 340 students) and a secondary school (with about 1,500 students) in the town of Fröndenberg and from the University of Duisburg-Essen (with about 37,000 students). All institutions are located in Germany's most populous state of North Rhine-Westphalia.[6]

In order to make the race game understandable to subjects of all age groups, we took great care in simplifying its exposition. After consulting several teachers, we opted for a purely graphical display of the games which was programmed in z-Tree (Fischbacher, 2007; see the screenshots in Figures 1 and 2 below). In addition, we used the following framing for the games: Subjects were informed that they are playing several games against a computer which "tries to win the game". They learned that the computer has hidden a treasure (the yellow square, see Figure 1) in a cave, but blocked the path from the cave's entrance to the treasure with stones (the red squares). The number of stones varies across games. In order to win the game, players have to reach the treasure by removing the stones. The subject and the

[5] As we only repeat every game once, improvements are more likely due to additional steps of reasoning rather than chance and reinforcement learning. The results by Gneezy et al. (2010) reveal that few subjects who have learned the winning strategy in $G(m=14, k=3)$ are subsequently able to win $G(m=16, k=4)$ on the first try.

[6] The details of the education system in Germany vary by state. Generally, after primary school, i.e. usually after grade 4, children can attend four types of secondary schools: Two types that offer degrees allowing to pursue different paths of vocational training (Hauptschule and Realschule), one type that aims at awarding the degree necessary for university admission (Gymnasium), and a fourth type that offers all types of degrees (Gesamtschule). Fröndenberg has only one secondary school which is of the latter type. Thus, selection effects through educational tracking are minimized. We further check for selection effects (if there are any) by recruiting two age groups from each type of school (and find significant age effects also within a school type) and run regressions including school marks (and find a significant effect of age also when including these controls).

computer take turns in removing stones by dragging them into their respective box which holds up to four stones. After each turn, the stones in the box disappear. Whoever is able to place the treasure in his box wins the game. In all twelve games subjects are the first mover and, accordingly, can always reach a winning position in the first move.[7] If the computer cannot reach a winning position, it resorts to random play.

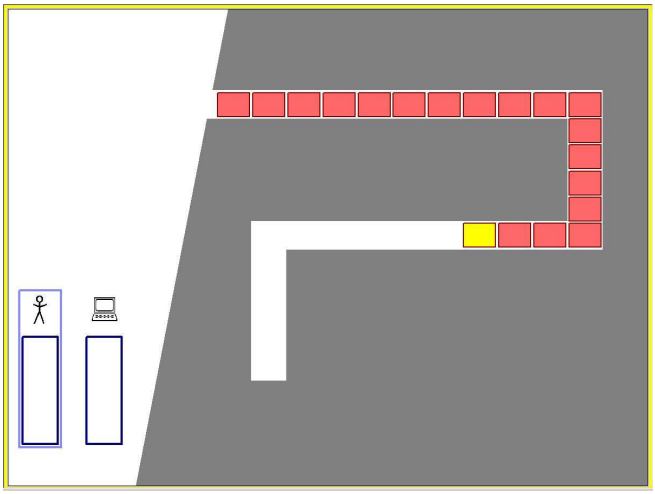


Figure 1: Graphical display of game $G(m=19, k=4)$

In order to gain insight into the information subjects acquire to solve the game and to identify chance winnings, we initially hide the length of the game (see Johnson et al., 2002, for a similar procedure). That is, subjects were informed that their view of the cave is blocked by bushes (the green squares, see Figure 2). In order to take a stone, the bushes covering it need to be removed by clicking on a pair of scissors. With each click, starting from the cave's entrance, two adjacent bushes disappear. During their turn subjects can remove as many bushes as they like. That is, subjects could remove bushes whenever they wanted during the game except for the time the computer made its moves.

[7] By playing against a computerized opponent, performance is comparable across all individual players (see, e.g., Johnson et al., 2002, McKinney and Van Huyck, 2006, Brosig and Reiß, 2007, and Burks et al., 2009, for similar approaches in sequential games). As the computer is programmed to win the game, it will reach the winning position as soon as the subject makes a wrong move. Accordingly, just imitating the moves made by the computer never pays off for subjects (even more so as optimal moves change with the length of the games). Nevertheless, we cannot exclude the possibility that observing the moves made by the computer after an own wrong move provides some help in learning to apply backward analysis.

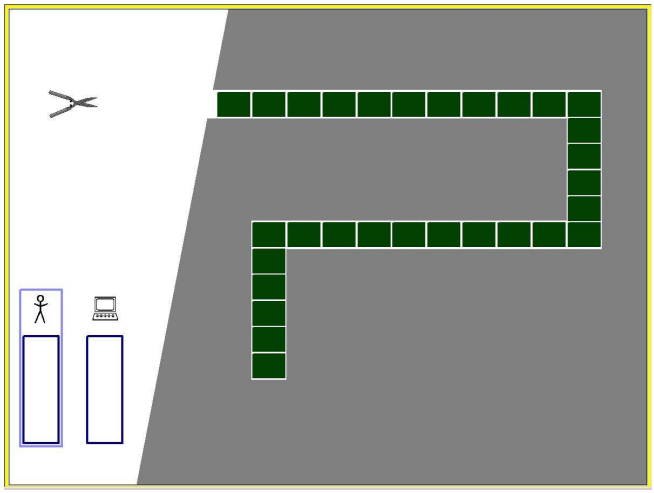


Figure 2: Graphical display of game with hidden length

At the beginning of the experiment, subjects received instructions, which were read aloud and accompanied by a presentation and a video.[8] After the presentation, five control questions were read aloud. Answers had to be given by dragging a ball into a "yes" or a "no" box. By using a similar elicitation procedure as in the games, we could also test whether subjects were able to handle the computer mouse (which was the case for all subjects). Having answered all questions correctly or having been taught the correct answers, subjects played the six race games $G_i(m, k)$ twice in the identical order with m taking the values 19, 3, 29, 8, 11, and 21 (but were left ignorant of the exact number of games to be played). The two series are indicated by the subscript i=1, 2. In the following we refer to the first series of the six games as part 1 and to the second series as part 2.

At the end of the experiment, 6th graders, 9th graders, and all university students had to fill out a questionnaire asking for personal characteristics such as risk attitudes and trust behavior. After filling out the questionnaire, subjects were paid off and received a fixed amount of money for each game won. We aimed at providing similar incentives for all subjects and therefore varied the amount across age groups. Students earned 5 Euro per game, 9th graders 4.40 Euro, 6th graders 2.70 Euro, 4th graders 1.80 Euro, and 1st graders 1 Euro.[9]

[8] All instructions and questionnaires are included in Online Appendices B and C. The video is available upon request. Before the first session, we tested the design with 7 children aged 8 to 13 who showed no problems in understanding the game.

[9] The incentives were set after consulting the school board of the respective schools. Furthermore, we based the calculation on public pocket money guidelines for children in North Rhine-Westphalia, which suggest monthly payments of 13.00 Euro for 7 year olds, 22.00 Euro for 9 and 10 year olds, 30.90 Euro for 12 year olds, and 48.50 Euro for 15 year olds (LWL, 2009). These guidelines are released by the union of municipalities of North Rhine-Westphalia and are intended, e.g., for public institutions that raise children (youth centers, etc.). As such, they are based on German social laws and determine how much pocket money is handed out to children and adolescents in those institutions.

In all age groups, subjects knew that their performance would be recorded anonymously (i.e., we used a double-blind procedure). Before the experiment, subjects received a card with a code name and were randomly assigned to a computer. At the end of the experiment subjects entered their code name and received their payment in a padded envelope marked with their code name from a person unaware of the amount it contained. At the schools the envelopes were handed out by teachers and at the university by a student assistant not involved in the experiment. At both schools it was necessary to collect written consent from the students' parents. To preserve anonymity, teachers collected the forms and were carefully instructed to randomly select eligible students from their class as subjects.

At both schools, we ran two sessions within each age group with 15 subjects each. All these sessions were conducted at a computer lab in the secondary school. At the university, we ran six sessions with a total of 55 university students. All these sessions took place at the Essen Laboratory for Experimental Economics (elfe). University students were recruited via Orsee (Greiner, 2004) such that the share of economics students among the subjects approximately matches the share of people from any given cohort who start studying economics in Germany. Our data set is summarized in Table 1.

Group	N^*	Female	Minimum Age	Maximum Age	Institution
Grade 1	30	67%	06 y 10 m	07 y 11 m	Elementary school
Grade 4	30	50%	09 y 10 m	11 y 08 m	Elementary school
Grade 6	30	47%	11 y 10 m	13 y 08 m	Secondary school
Grade 9	30	43%	14 y 11 m	16 y 11 m	Secondary school
University	55	51%	20 y 00 m	26 y 01 m	University

* As there is no interaction between subjects in the experiment, *N* denotes the number of subjects as well as the number of statistically independent observations.

Table 1: Age groups

5 Results

For the analysis, we homogenized data sets among age groups. That is, within each age group we selected the largest group within a common age range of 12 months.[10] This served to avoid potential biases in the results due to students repeating grades, for example. The resulting data set (including age ranges) is summarized in Table 2. In the following three subsections we present descriptive statistics and non-parametric tests on the three hypotheses derived in section 2. The fourth subsection provides further evidence from regression

[10] The threshold age between older and younger university students is set at the median age within this group. Data at the individual level is included in Online Appendix D.

analyses. Note that using the full sample does not alter our main results (regressions based on the non-homogenized data set are included in Online Appendix E and referred to in section 5.4).

Group	*N*	Female	Minimum age	Maximum age
Grade 1	29	66%	06 y 10 m	07 y 09 m
Grade 4	28	50%	09 y 11 m	10 y 10 m
Grade 6	25	44%	11 y 10 m	12 y 09 m
Grade 9	24	38%	14 y 11 m	15 y 10 m
University young	21	57%	20 y 00 m	20 y 10 m
University old	19	53%	23 y 00 m	23 y 11 m

Table 2: Restricted data set

5.1 The ability to reason backwards

As Hypothesis 1 relates to the initial ability to reason backwards, we focus on the first series of the six race games (i.e., part 1) when testing this hypothesis. Figure 3 illustrates the average number of games won by the different age groups in this part. On average, subjects win 1.897 of the six games (and, not surprisingly, they are more likely to win a short game than a long game, see Figure 4). Although there appear to be differences between age groups (see section 5.3), in all groups the 95-percent confidence interval for the number of games won starts at, at least, 1.075 games (1st graders) and ends at, at most, 2.829 games (old university students).[11]

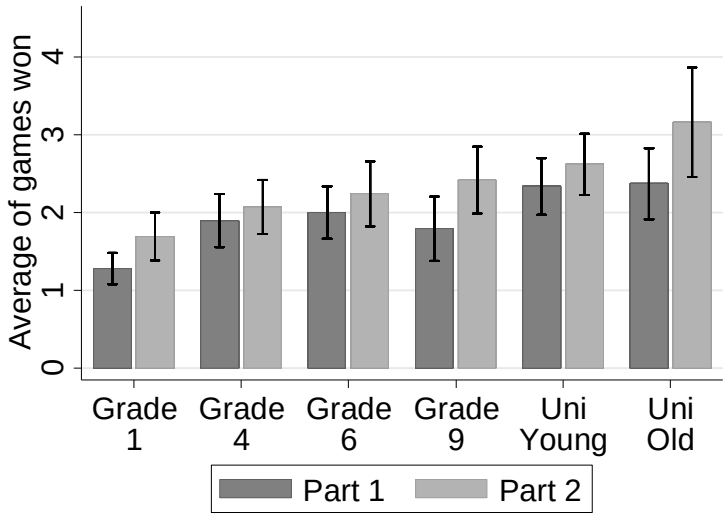


Figure 3: Average number of games won by age groups (with 95% confidence intervals)

[11] One possible explanation for the rather poor performance of 1st graders in part 1 might be that they are too young to understand the instructions. However, all except two 1st graders won the shortest game $G_1(m=3, k=4)$ already in part 1. Excluding these two subjects from the data set does not change any results on the performance qualitatively.

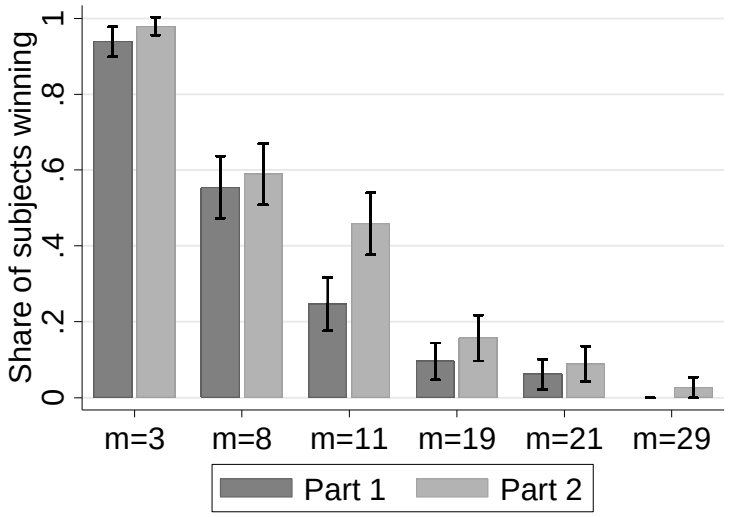


Figure 4: Share of subjects winning by game length (with 95% confidence intervals)

One necessary (though, not sufficient) condition indicating that subjects reason backwards is that they uncover the length of the game (i.e., remove the bushes in order to uncover the treasure) before their first move. Overall, 62.3 percent of subjects *always* uncover the treasure before playing a game in part 1. This share is largely driven by the first game of this part in which the treasure is uncovered by 65.1 percent of subjects. The percentage of subjects *always* uncovering the treasure varies across age groups from 41.7 percent for 6th graders to 78.9 percent for old university students. In subsequent games this share never drops below 75.0 percent for any of the six age groups, suggesting that some learning takes place already within part 1. Moreover, subjects who win a game almost always uncover the treasure beforehand. Out of the 277 won games only 1.4 percent are won by a player who does not uncover the length of the game. In contrast, in 21.0 percent of the 599 lost games the length of the game is not uncovered before the first move. That those who do not uncover the treasure before their first move are more likely to lose also holds within all of the six age groups.

Since uncovering the length of the game does not necessarily imply that subjects subsequently reason backwards, we also test whether those who uncovered the treasure perform significantly better than chance. In particular, we calculate the probability of chance winnings (based on the programmed computer play) for each game in part 1. Comparing observed frequencies of winning with the probabilities of chance winnings we find that subjects on aggregate perform significantly better than chance in all games except the longest game (i.e., game $G_1(m=29,\ k=4)$; $p<0.001$, one-tailed exact binomial tests[12]). This is also true when

[12] We use one-tailed exact binomial tests for comparing observed frequencies of winning to expected frequencies.

differentiating between age groups (p<0.039, one-tailed exact binomial tests), except for young university students (who do not solve the second longest game significantly better than chance), old university students (who do not solve the third longest game significantly better than chance), and 1st graders (who only solve the two shortest games and the third longest game significantly better than chance; p<0.080)[13]. Similar results apply if the calculated probability is based on the additional assumption that the treasure is taken as soon as it is within reach, i.e. that on the last move the winning strategy is selected with probability one (p<0.001 for all subjects, p<0.069 for all age groups except the second longest game for young university students, the third longest game for 9th graders and old university students, and all games for 1st graders in which p>0.145; one-tailed exact binomial tests).[14]

Our findings regarding Hypothesis 1 can be summarized as follows: First (and in line with previous results), subjects have difficulties in solving the race games. This is true for all age groups and particularly pronounced for 1st graders. Second, our findings on subjects' information acquisition and their performance once the length of a game is known provide some support for the hypothesis that subjects of all age groups are able to reason backwards, at least to some extent (though, the evidence for 1st graders is rather weak).

5.2 Improvement of the ability to reason backwards

Comparing subjects' performance between the two series of race games, we observe better performance in the second series for all age groups, significantly so for 1st graders, 9th graders, and old university students (p<0.028, two-tailed Wilcoxon signed rank tests[15]; see Figure 3). Table 3 summarizes the change in performance from part 1 to part 2. In all age groups, the majority of subjects wins at least as many games in part 2 as they did in part 1. As a result, the number of games won on average increases from 1.897 in part 1 to 2.301 in part 2. Performance increases (or, for a minority of subjects, decreases) on average by between one and two games. Of course, these increases do not capture any learning that happens within part 1 or part 2.

[13] Note that the number of 1st graders solving one of the longer games is rather low (i.e., it is never higher than 10.3 percent). See Online Appendix F for a graphical display of the share of subjects winning by game length and age group.

[14] Even if subjects are assumed to solve short games with $m \leq 8$ with certainty and play randomly otherwise, i.e. get the last two moves right, observed winning frequencies exceed the calculated frequencies significantly in all games, but $G_1(m=11, k=4)$ and $G_1(m=29, k=4)$ (p<0.074). Note, however, that under such additional assumptions the short games are not testable anymore as these games would be won with certainty (which is not the case either, see Figure 4).

[15] We use two-tailed Wilcoxon signed rank tests for non-parametric within-subject comparisons based on (at least) ordinal data.

Share of subjects (in percent)																		
Grade 1			Grade 4			Grade 6			Grade 9			Uni Young			Uni Old		
+	=	-	+	=	-	+	=	-	+	=	-	+	=	-	+	=	-
48.3	34.5	17.2	35.7	46.4	17.9	40.0	32.0	28.0	58.3	33.3	8.3	38.1	47.6	14.3	52.6	31.6	15.8

Table 3: Change in performance between parts 1 and 2

Looking at each of the six games separately, we find that the share of subjects solving a game significantly increases for all games except $G(m=8, k=4)$ ($p<0.05$, two-tailed exact McNemar test[16]; see Figure 4). The increase is particularly pronounced for $G(m=11, k=4)$. However, even when we exclude this game from our analysis, we observe that 1st graders, 9th graders, and old university students solve (weakly) significantly more games in part 2 than in part 1 (1st graders: $p<0.050$, 9th graders and old university students: $p<0.100$, two-tailed Wilcoxon signed rank tests).[17]

Do the improvements in performance observed in part 2 imply that subjects are better able to reason backwards? We find that the share of subjects always uncovering the treasure before playing a game significantly increases from 62.3 percent in part 1 to 82.2 percent in part 2 ($p=0.000$; two-tailed exact McNemar test). This result holds for all age groups except 1st graders ($p<0.083$, 1st graders $p=0.655$; two-tailed exact McNemar tests). Note that this effect can be mainly attributed to the rather low share of subjects always uncovering the treasure in the first game in part 1. When excluding this game from the analysis the increase is no longer significant (i.e., 78.1 percent in part 1 versus 82.2 percent in part 2; $p=0.157$ overall and $p>0.157$ for each of the six age groups; two-tailed exact McNemar tests). This result suggests that a significant part of learning to uncover the game already occurs in part 1. Nevertheless, when referring to the average sum of games in which a subject uncovers the treasure before playing this game, we observe significant improvements between the two parts also when excluding the first game in each of the parts ($p=0.000$ including, $p=0.017$ excluding the first games, two-tailed Wilcoxon signed rank tests). When testing separately for each age group, this result only holds for 4th graders, however ($p=0.087$ for 4th graders, $p>0.156$ for all other age groups when excluding the first game, two-tailed Wilcoxon signed rank tests).

[16] We use two-tailed exact McNemar tests for non-parametric within-subject comparisons based on nominal data.

[17] We also run ordered probit regressions with a dependent variable indicating the change of performance (improved/stayed constant/deteriorated) for each of the games separately (except the longest and the shortest games). The results suggests that the increase in performance for game $G(m=19, k=4)$ is particularly driven by 1st graders while the increase in performance for game $G(m=21, k=4)$ is particularly driven by young and old university students. The regression is included in Online Appendix E.1.

Note that also in part 2 we observe that subjects who win a game almost always uncover the treasure beforehand. Out of the 336 won games in part 2 only 1.8 percent are won by a player who does not uncover the length of the game. In contrast, in 13.5 percent of the 540 lost games the length of the game is not uncovered before the first move. Again, also within all of the age groups it holds that those who do not uncover the treasure before their first move are more likely to lose.

The significant increase in performance in part 2 also displays in our tests on whether those who uncovered the treasure perform significantly better than chance. In particular, in part 2 subjects perform significantly better than chance in all six games ($p<0.001$, one-tailed exact binomial tests). In contrast to part 1, in the second part 1st graders also solve game $G_2(m=11, k=4)$ significantly better than chance, 4th graders and old university students also solve the longest game significantly better than chance, young university students also solve the second longest game significantly better than chance, and old university students also solve the third longest game significantly better than chance. Subjects still perform significantly better than chance in all six games if the calculated probability is based on the additional assumption that subjects take the treasure as soon as it is within reach, i.e. get the last move right ($p<0.001$, one-tailed exact binomial tests). Compared to part 1, now 1st graders also solve games $G_2(m=11, k=4)$ and $G_2(m=19, k=4)$ significantly better than chance, 4th graders and old university students also solve the longest game significantly better than chance, young university students also solve the second longest game significantly better than chance, and 9th graders and old university students also solve the third longest game significantly better than chance.[18]

To summarize, our findings only partly support Hypothesis 2. While performance generally improves with repetition, it is significant for 1st graders, 9th graders, and old university students, only. Our analysis of subjects' information acquisition and their performance once the length of a game is uncovered is in line with the interpretation that the increase in performance is related to an enhanced capability to reason backwards.

We also find some indication for improvement of backward reasoning within parts when comparing the number of shorter games that are won after winning or losing a specific game.

[18] Note that with regard to both specifications, 4th graders no longer solve the second longest game and young university students no longer solve the third longest game significantly better than chance in part 2. Importantly, even if subjects are assumed to select the successful strategy with probability one once the treasure comes within reach in the *next* move, observed winning frequencies exceed the calculated frequencies significantly in all six games in part 2 ($p<0.003$). Under such additional assumptions the short games with $m\leq 8$ are not testable anymore as these games would be won with certainty (which is not the case either, see Figure 4).

The analysis focuses on games $G_1(m=19, k=4)$, $G_1(m=11, k=4)$, $G_1(m=21, k=4)$, $G_2(m=19, k=4)$, and $G_2(m=29, k=4)$ (the other games have to be excluded either because the number of winners and losers is strongly unbalanced in these or subsequent shorter games, or there is no shorter game afterwards). We observe that for all five games subjects who win this game solve more shorter games afterwards than subjects who lose this game. The difference is significant for games $G_1(m=19, k=4)$, $G_1(m=11, k=4)$, and $G_2(m=29, k=4)$ ($p<0.071$, two-tailed Mann-Whitney-*U* test[19]). Running this test for each age group separately, we find similar results. Note that the latter analysis is based on game $G_1(m=11, k=4)$ only as in this game the number of observations for winners and losers is reasonably balanced in most age groups (it is still very unbalanced for 1st graders and 6th graders). We find that winners solve more subsequent shorter games than losers and that his difference is significant for young and old university students ($p<0.061$, two-tailed Mann-Whitney-*U* tests).

5.3 The effects of age and gender

Looking for differences between age groups, we find that 1st graders perform significantly worse than all other age groups in part 1 ($p<0.016$, two-tailed Mann-Whitney-*U* tests). Moreover, in part 1 both groups of university students perform weakly significantly better than 4th and 9th graders ($0.057<p<0.089$). All other differences between age groups are not significant ($p>0.115$) in this part. Our results included in the last section reveal that those who, in part 1, are significantly less able to solve the games than the subsequent age group (i.e., 1st graders and 9th graders) significantly improved their performance in part 2. As a result, in part 2 1st graders no longer perform significantly worse than 4th graders ($p=0.124$, two-tailed Mann-Whitney-*U* test), but still significantly worse than all other age groups ($p<0.052$), and 9th graders no longer perform significantly worse than young university students ($p=0.548$), but still (weakly) significantly worse than old university students ($p=0.067$).

[19] We use two-tailed Mann-Whitney-*U* tests for non-parametric between-subject comparisons based on (at least) ordinal data. Following Dufwenberg et al. (2010) and defining the moment after which no mistakes are made anymore as the moment of full epiphany, we find that only a few subjects seem to reach this full epiphany in our experiment, however. As shown in Figure 4, the very last game $G_2(m=21, k=4)$ is solved by only 13 subjects. Of those, 1 subject makes the last mistake in the eleventh game $G_2(m=11, k=4)$, 6 subjects in the ninth game $G_2(m=29, k=4)$, 3 subjects in the seventh game $G_2(m=19, k=4)$, and 1 subject in the sixth game $G_1(m=21, k=4)$. One possible reason for this rather low rate of complete epiphany is that subjects in our experiment never play the same game twice in a row and mistakes of play are immediately punished by the computerized opponent playing the optimal strategy. This is also in line with a recent study of the race game conducted by Hawes et al. (2012). They report that subjects improve their performance incrementally rather than developing the optimal solution at once and support this finding by both behavioral and fMRI data.

Considering all age groups collectively, male subjects perform significantly better than female subjects in both series of games (p<0.011, two-tailed Mann-Whitney-U tests). Differentiating between age groups and the two series reveals that the significant differences are restricted to both 4th graders and 6th graders in the second series of race games (p<0.027). Moreover, comparing the performance of subjects across the two parts of the experiment, we find significant improvements for male 4th graders and female 9th graders as well as for female old students (p<0.047, two-tailed Wilcoxon signed rank tests). We find weakly significant improvements for female 1st graders, male 9th graders, and male young students (0.083<p<0.098). The data is illustrated in Figure 5.

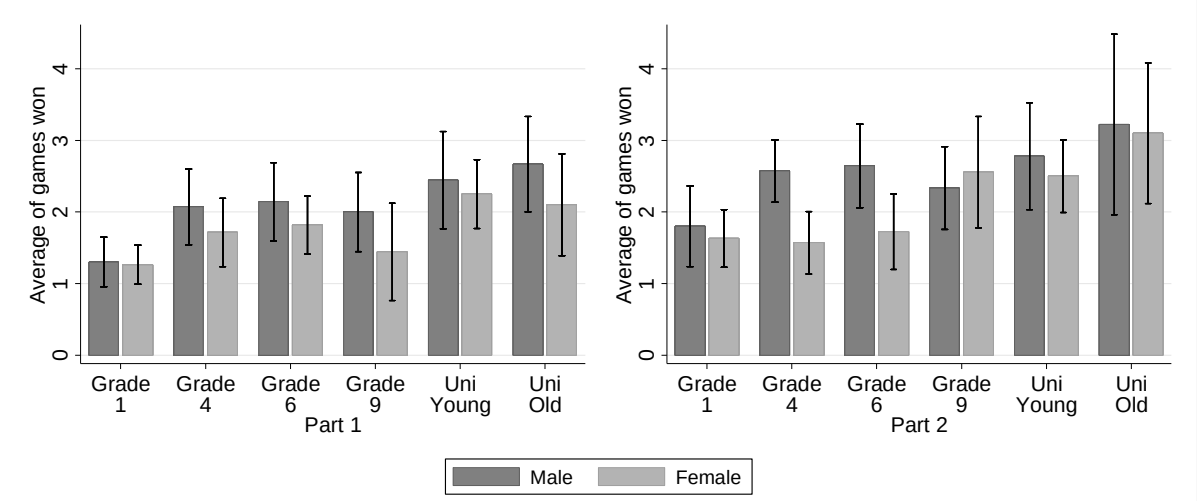


Figure 5: Average number of games won by male and female subjects (with 95% confidence intervals)

These results suggest differences in the development of the capability to solve the race game and to reason backwards between males and females: In line with Hypothesis 3, the ability to reason backwards significantly improves from 1st graders to old university students for both genders (p<0.011, two-tailed Mann-Whitney-U tests). Moreover, male and female subjects in our sample start off and end up at the same level of performance (p>0.453, two-tailed Mann-Whitney-U tests). But males seem to acquire some skills that help to improve performance in the game at earlier ages than females. The main development step appears to take place sometime between grades 1 and 4 for males and sometime between grades 6 and 9 for females.

Our findings are similar to the ones obtained in the study on the development of planning behavior conducted by Bishop et al. (2001). Also with respect to Tower of Hanoi tasks gender-related performance differences are restricted to 9 to 12 year olds (see section 2). Moreover, we re-investigate data obtained by Harbaugh et al. (2001), which focuses on

rational choice behavior, but does not present evidence on gender effects. Our analysis reveals similar age-related differences between males and females in their study. In particular, we find that there are no significant gender differences in the number of inconsistent decisions for 2nd graders and university students ($p>0.500$, two-tailed Mann-Whitney-U tests), but that male 6th graders decide more consistently than their female peers ($p=0.020$).

5.4 Further analyses

Next to age and gender we collected further information about the individuals participating in our study. In order to control for potentially confounding influences, we additionally estimate several regression models that include this information. We conduct three analyses using different dependent variables: We analyze (i) the number of games won within part 1 and within part 2, (ii) the difference in number of games won between part 1 and part 2, and (iii) the number of correctly identified winning positions during the play of the games within part 1 and within part 2. We conduct these analyses also for the full data set without homogenizing age groups as well as for both data sets (with and without homogenized age groups) excluding the shortest game $G(m=3, k=4)$. The results of these robustness checks are included in Appendices E.2 to E.8 and are referred to in this section.

In analysis (i) we present OLS-regressions using the number of games won by a subject as dependent variable separately for both parts. Specification (1) includes dummy variables for grades, for gender, and for whether the subject uncovered the treasure before the very first move in the first of the 12 games played in total. Specification (2) additionally includes school marks received in math and in German. It is based on a data set that excludes all 1st graders, since they do not receive marks yet. Specification (3) additionally includes data from our ex-post questionnaire, i.e. data on a subject's patience, trust, fairness, and risk preferences. The data were elicited through questions similar to those employed in the German Socio-Economic Panel (a detailed description of variables is included in Appendix A). Since questionnaires were not completed by 1st graders and 4th graders, the data set for the third specification is restricted to 6th graders, 9th graders, and university students. In all regressions 6th graders serve as the baseline category. The results of the regressions are displayed in Tables 4 and 5.[20]

[20] Through the grade dummies our specifications capture non-linear age effects that are likely to occur especially between school students and university students due to selection. The variable 'uncover treasure' is included as a control for the initial understanding of the game. After the first game, the uncovering behavior does not exhibit much variation. For pupils the marks for German and math were obtained from the school directly, while university students were asked about their last marks in the questionnaire. After consulting teachers at both schools grades were aligned to a common scale from 1 (poor) to 15 (very good). Furthermore, as no participant wins every game or no game at all, censoring does not need to be taken into account.

The three regression specifications are run separately for the two parts of the experiment and support our previous interpretation of the data. Uncovering the location of the treasure before the very first move positively influences the number of games won in both parts, while being female has a (weakly) significantly negative effect on performance in 4th and 6th grade in part 2. When excluding the shortest game $G_1(m=3, k=4)$ we additionally observe a weakly significantly negative gender effect for female old university students in part 1 for specifications (1) and (2). Without homogenizing age groups there is also a (weakly) significantly negative effect of being female for 9th graders across the three specifications. These differences disappear in part 2 (see the Online Appendices E.2 and E.3).

	OLS - dependent variable: number of games won in part 1					
	(1)		(2)		(3)	
Grade 1	-0.966***	(0.331)				
Grade 4	-0.071	(0.300)	-0.130	(0.313)		
Grade 9	-0.058	(0.296)	-0.026	(0.309)	-0.040	(0.322)
Uni young	0.184	(0.341)	0.107	(0.357)	0.088	(0.392)
Uni old	0.459	(0.340)	0.431	(0.355)	0.395	(0.389)
Grade 1 x female	0.092	(0.312)				
Grade 4 x female	-0.426	(0.301)	-0.277	(0.319)		
Grade 6 x female	-0.234	(0.321)	-0.133	(0.344)	-0.114	(0.364)
Grade 9 x female	-0.492	(0.335)	-0.461	(0.349)	-0.512	(0.366)
Uni young x female	-0.048	(0.353)	-0.049	(0.370)	-0.035	(0.399)
Uni old x female	-0.577	(0.365)	-0.607	(0.385)	-0.589	(0.418)
Uncover treasure	0.480***	(0.146)	0.529***	(0.170)	0.510**	(0.202)
Mark German			-0.020	(0.042)	-0.018	(0.049)
Mark math			0.098***	(0.034)	0.102**	(0.040)
Fairness					-0.118	(0.224)
Patience					-0.018	(0.040)
Risk					0.151	(0.173)
Trust					0.004	(0.046)
Constant	1.835***	(0.232)	1.070**	(0.489)	0.761	(0.882)
Adj. R-squared	0.209		0.152		0.106	
N	146		116		88	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table 4: Regressions on the number of games won in part 1

	OLS - dependent variable: number of games won in part 2					
	(1)		(2)		(3)	
Grade 1	-0.932**	(0.404)				
Grade 4	-0.071	(0.367)	-0.125	(0.377)		
Grade 9	-0.248	(0.362)	-0.213	(0.372)	-0.225	(0.405)
Uni young	0.049	(0.417)	-0.027	(0.431)	0.006	(0.494)
Uni old	0.532	(0.415)	0.505	(0.428)	0.499	(0.491)
Grade 1 x female	-0.075	(0.382)				
Grade 4 x female	-1.050***	(0.368)	-0.917**	(0.384)		
Grade 6 x female	-0.850**	(0.392)	-0.850**	(0.414)	-0.860*	(0.459)
Grade 9 x female	0.269	(0.41)	0.300	(0.421)	0.165	(0.462)
Uni young x female	-0.171	(0.431)	-0.160	(0.446)	-0.113	(0.503)
Uni old x female	-0.130	(0.446)	-0.153	(0.464)	-0.128	(0.527)
Uncover treasure	0.349*	(0.179)	0.425**	(0.204)	0.463*	(0.255)
Mark German			-0.021	(0.050)	-0.016	(0.062)
Mark math			0.090**	(0.041)	0.093*	(0.050)
Fairness					-0.276	(0.282)
Patience					-0.004	(0.050)
Risk					0.419	(0.218)
Trust					0.023*	(0.058)
Constant	2.419***	(0.284)	1.727***	(0.589)	0.542	(1.111)
Adj. R-squared	0.200		0.178		0.097	
N	146		116		88	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table 5: Regressions on the number of games won in part 2

Controlling for school marks in specification (2) we additionally find that subjects who are better in math are also more capable of solving the games and, accordingly, win more games in the two parts. The inclusion of subjects' answers to the ex-post questionnaire in specification (3) reveals a weakly significantly positive relationship between trust and a subject's ability to solve the games in part 2. The effects of the mark in math and self-reported trust are robust to the choice of the data set. Also when pooling observations from part 1 and part 2 we observe similar effects of the mark in math and of self-reported trust (see Table E.4.1 in the Online Appendix E.4). This somewhat surprising result is in line with previous research by Burks et al. (2009) who find a positive correlation between trust and IQ. Neither risk attitudes nor fairness preferences appear to be related to the number of games won. Applying ordered probit regressions instead of OLS yields similar results for part 1, part 2, and when pooling both parts (see Tables E.2.2, E.3.2, and E.4.2 in the respective Online Appendices).

Applying Wald tests on regression specification (1) yields more insights about gender-specific age differences. The results support our conclusions drawn in the previous section and are in line with Hypothesis 3: Performance improves with age and males seem to acquire some skills that help to win the game at earlier ages than females. In particular, male 1st graders win (weakly) significantly less often than males of all older age groups in both parts (p<0.036 for nine out of ten, p=0.092 for one comparison), while there is almost no significant difference between the older age groups (p>0.120 for 19 out of 20, p=0.061 for one comparison). In contrast, in both parts the performance of females is almost never significantly different between 1st, 4th, and 6th graders (p>0.180 for five out of six, p=0.036 for one comparison), but females of these three younger age groups win (weakly) significantly less often than females of the three older age groups (p<0.049 for eleven out of eighteen, p=0.075 for one, p>0.250 for six comparisons). Again, between the three older age groups there is almost no significant difference regarding the performance (p<0.210 for five out of six, p=0.053 for one comparison).

After analyzing the performance within the two parts we now turn to the differences in performance between part 1 and part 2 and their relationship with age and gender. Analysis (ii) which includes the difference in number of games won between part 1 and part 2 as a dependent variable provides some insights into this relationship. In this analysis we use the same independent variables as above yielding three similar specifications. The results of the respective OLS-regressions are presented in Table 6. Ordered probit regressions produce similar results (see Table E.5.2 in the Online Appendix E.5).

For specifications (1) and (2) it shows that performance of female 4th graders improves weakly significantly *less* than performance of male 4th graders while performance of female 9th graders improves weakly significantly *more* than performance of male 9th graders. This observation on relative performance increases extends the results on improvement in section 5.3 for these age groups. However, the effect for male 9th graders turns insignificant after including additional controls in specification (3). Wald tests comparing the improvement of performance between different age groups for the same gender using specification (1) reveal no significant differences for males. Yet, they reveal that the performance of female 9th graders and that of female old university students increases significantly more than that of female 4th and 6th graders (p<0.010). It also increases weakly significantly more than that of female 1st graders and female young university students (p<0.100).

	OLS - dependent variable: difference in number of games won between part 1 and part 2					
	(1)		(2)		(3)	
Grade 1	0.034	(0.407)				
Grade 4	0.000	(0.369)	0.006	(0.378)		
Grade 9	-0.190	(0.364)	-0.188	(0.373)	-0.185	(0.417)
Uni young	-0.134	(0.419)	-0.133	(0.432)	-0.081	(0.509)
Uni old	0.073	(0.418)	0.074	(0.429)	0.104	(0.505)
Grade 1 x female	-0.167	(0.384)				
Grade 4 x female	-0.624*	(0.370)	-0.641*	(0.385)		
Grade 6 x female	-0.616	(0.395)	-0.717*	(0.415)	-0.746	(0.473)
Grade 9 x female	0.760*	(0.412)	0.761*	(0.422)	0.678	(0.475)
Uni young x female	-0.123	(0.434)	-0.112	(0.447)	-0.078	(0.518)
Uni old x female	0.447	(0.449)	0.453	(0.465)	0.460	(0.542)
Uncover treasure	-0.131	(0.180)	-0.104	(0.205)	-0.047	(0.262)
Mark German			-0.001	(0.050)	0.002	(0.064)
Mark math			-0.009	(0.041)	-0.009	(0.052)
Fairness					-0.158	(0.291)
Patience					0.014	(0.052)
Risk					0.019	(0.059)
Trust					0.268	(0.225)
Constant	0.584**	(0.285)	0.657	(0.591)	-0.219	(1.144)
Adj. R-squared	0.035		0.042		-0.026	
N	146		116		88	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table 6: Regressions on improvement between part 1 and part 2

In analysis (iii) we additionally conduct a more fine-grained investigation based on each of the moves the subjects made in the experiment. The dependent variable in this analysis is the number of correctly identified winning positions during the play of the games. A game $G(m, k)$ is made up of m positions of which a subset are winning positions. Each move a subject makes in the experiment can lead to a winning or a losing position. A winning position can be characterized by the distance to the treasure n. More precisely, n indicates how many correct moves are necessary to reach the treasure from this point in the game (cf. section 3). The following analysis considers whether a subject has uncovered the treasure on a move-by-move basis. That means, we only count a move as "correct" if it was made *after* uncovering the treasure and if it resulted in a winning position. Moves to a winning position that were made by luck, i.e. without knowing the position of the treasure, are considered to be wrong as they cannot be regarded as evidence for any sort of backward analysis.

Figure 6 shows the error rates over the positions n across grades and gender aggregated over part 1 and part 2. These error rates are calculated conditional on reaching the respective

position. Note that the number of observations available for calculating the error rates decreases with *n*. The longest game $G(m=29, k=4)$ potentially provides us with observations for $n=6$ to $n=1$ while the shortest game $G(m=3, k=4)$ only yields one observation in $n=1$. The figure reveals, not surprisingly, that error rates roughly increase with the distance from the treasure. However, it mainly highlights the gender differences already suggested by the regressions above as well as by the tests in the previous section. There is a pronounced difference in 4th and 6th grade: In these age groups female participants never have a lower error rate than their male peers. In all other age groups the difference in error rates is smaller and female players exceed the performance of males in at least one of the positions *n*.

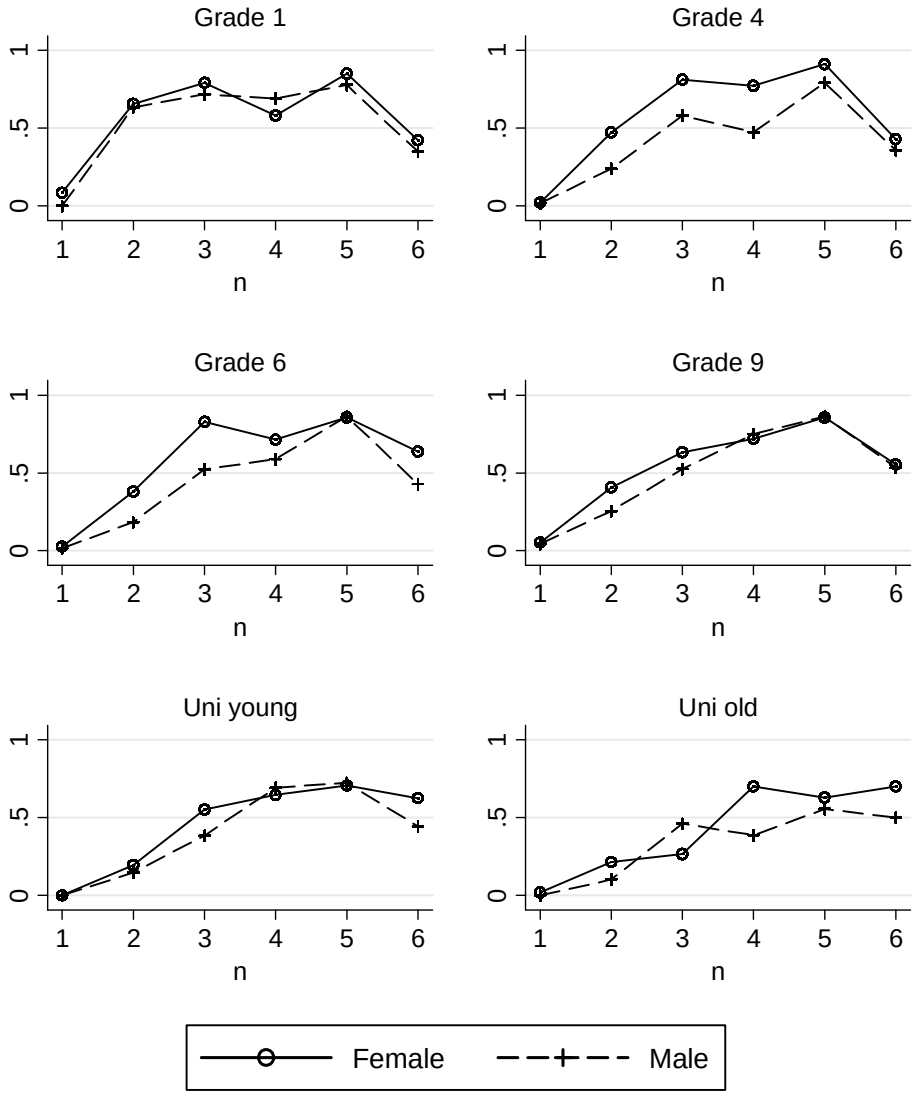


Figure 6: Error rates conditional on reaching position *n*

In most age groups there is a drop in error rates in $n=6$. This effect is driven only by the longest game, namely $G(m=29, k=4)$. The winning strategy in this game requires subjects to

take four stones in the first move. This move coincides with the commonly observed behavior to "shorten" the game as much as possible. In part 1, 58 percent of the subjects make a correct move in n=6 by taking four stones, but none of these subjects is able to find the next winning position in n=5 that requires them to take three stones. In part 2, 44 percent make the correct move in n=6. 19 percent of them also find the next winning position in n=5. Overall only four subjects win this game.

Analyzing the performance depending on the position of the game more formally, we assume that each subject plans his or her action prior to every move i in some way that depends on n. Each move made by subject i is characterized by δ_i indicating whether it leads to a winning position (δ_i=1) or a losing position (δ_i=0). The ability to make a correct move is determined by the positive random variable T that can be thought of as the number of correctly identified winning positions in the remaining game or the successful steps of backward reasoning. As long as a draw t of T is larger or equal to n, the move will be correct and the subsequent winning position will be reached. For simplicity, we assume T to be continuous and having a conditional density $f(t|x,\theta)$ where x is a vector of subject specific covariates and θ is a parameter vector. Then, building on the techniques of survival analysis, we refer to the cumulative distribution function of T as $F(t|x,\theta)=\Pr(T<t|x,\theta)$ and to the complementing survival function as $S(t|x,\theta)=\Pr(T\geq t|x,\theta)=1-F(t|x,\theta)$ (see, e.g., Klein and Moeschberger, 2003).

In the experiment we never observe the draws t directly. Instead our observations are censored. If we observe a wrong move in position n, we only know that an error of reasoning was made, but we do not learn the realization of t, i.e. $0 < t < n$. If we observe a correct move in n then we know that no error occurred in n or before. However, we do not know how many correct moves could have followed, i.e. $n \leq t$. This yields a likelihood function

$$L(\theta) = \prod_i S(t_i|x,\theta)^{\delta} \, (1-S(t_i|x,\theta))^{1-\delta}$$

over all moves i. In order to maximize this likelihood function, however, some assumption over the distribution of T has to be made. We selected the generalized gamma distribution from a set of commonly used distributions based on the Bayesian information criterion.[21]

[21] We also considered the exponential, the Weibull, the Gompertz, the log-normal, the log-logistic, the inverse Gaussian and the gamma distribution (see Tables E.6.1 and E.7.1 in the respective Online Appendices). Regressions were performed with Stata's user-written command "intcens" for interval censored data (Griffin, 2005). In addition, we considered two alternative approaches that yielded similar results: First, the method introduced by Royston and Parmar (2002) that models the baseline cumulative hazard function of a proportional hazards model as a cubic spline. Second, a Cox proportional hazards model with right-censoring and midpoint imputation. Regressions based on the former approach are included in Tables E.6.4 and E.7.4 of the respective Online Appendices. Regressions based on the latter approach are included in Tables E.6.5 and E.7.5.

We run three different models with the independent variables also used in analyses (i) and (ii) above. We report the exponentiated coefficients of the accelerated failure time models in Tables 7 and 8.

	Survival analysis - dependent variable: correctly identified winning positions					
	(1)		(2)		(3)	
Grade 1	0.841	(0.146)				
Grade 4	0.988	(0.14)	0.947	(0.145)		
Grade 9	0.884	(0.142)	0.848	(0.145)	0.871	(0.149)
Uni young	1.306	(0.283)	1.214	(0.242)	1.255	(0.302)
Uni old	1.399*	(0.272)	1.308	(0.251)	1.369	(0.341)
Grade 1 x female	0.845	(0.102)				
Grade 4 x female	0.863	(0.140)	0.881	(0.148)		
Grade 6 x female	0.862	(0.146)	0.922	(0.153)	0.957	(0.206)
Grade 9 x female	0.798	(0.157)	0.805	(0.155)	0.808	(0.154)
Uni young x female	0.863	(0.158)	0.884	(0.151)	0.901	(0.162)
Uni old x female	0.731**	(0.109)	0.762*	(0.118)	0.740*	(0.121)
Mark German			0.985	(0.018)	0.992	(0.020)
Mark math			1.028	(0.020)	1.024	(0.020)
Fairness					0.974	(0.089)
Patience					0.994	(0.017)
Risk					0.996	(0.020)
Trust					1.082	(0.081)
Constant	1.457*	(0.31)	1.441	(0.340)	1.099	(0.529)
Log pseudolikelihood	-694.365		-551.088		-414.951	
Number of moves	1,317		1,073		821	
Number of individuals	146		116		88	

Robust standard errors clustered on the individual level are given in parentheses. Coefficients of the accelerated failure time model are exponentiated. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table 7: Regressions on correctly identified winning positions in part 1

	Survival analysis - dependent variable: correctly identified winning positions					
	(1)		(2)		(3)	
Grade 1	0.719*	(0.132)				
Grade 4	0.999	(0.137)	0.997	(0.128)		
Grade 9	0.835	(0.147)	0.855	(0.151)	0.860	(0.138)
Uni young	1.090	(0.201)	1.071	(0.183)	1.078	(0.179)
Uni old	1.292	(0.262)	1.285	(0.264)	1.273	(0.259)
Grade 1 x female	0.851	(0.162)				
Grade 4 x female	0.643***	(0.082)	0.670***	(0.082)		
Grade 6 x female	0.645***	(0.096)	0.650***	(0.096)	0.657***	(0.093)
Grade 9 x female	1.008	(0.191)	1.022	(0.185)	0.935	(0.157)
Uni young x female	0.897	(0.160)	0.894	(0.147)	0.933	(0.148)
Uni old x female	0.807	(0.220)	0.856	(0.207)	0.950	(0.212)
Mark German			0.996	(0.021)	0.994	(0.023)
Mark math			1.026	(0.019)	1.026	(0.020)
Fairness					0.917	(0.092)
Patience					1.006	(0.020)
Risk					1.026	(0.020)
Trust					1.180**	(0.090)
Constant	2.317***	(0.305)	2.090***	(0.569)	1.291	(0.545)
Log pseudolikelihood	-694.324		-521.716		-389.235	
Number of moves	1,424		1,136		880	
Number of individuals	146		116		88	

Robust standard errors clustered on the individual level are given in parentheses. Coefficients of the accelerated failure time model are exponentiated. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table 8: Regressions on correctly identified winning positions in part 2

The regressions reveal no or only weakly significant differences in performance during part 1. Old female university students perform slightly worse than their male counterparts ($p\leq0.080$) across specifications. Old university students tend to perform somewhat better than 6th graders as long as we do not control for grades or fairness, patience, risk, or trust attitudes ($p=0.085$). The gender differences described above appear to be mainly driven by the performance in the second part. In this part, female 4th and 6th graders perform significantly worse than their male class mates: in 4th grade the distance from the treasure in which females successfully solve the game is on average 35.8 percent smaller than that of males ($p\leq0.001$); in 6th grade it is 35.5 percent smaller ($p\leq0.004$).[22] This significant effect is robust across specifications and provides further evidence for gender-specific age effects as formulated in Hypothesis 3. We also find a significant increase of performance associated with larger trust ($p=0.030$) and a weakly significantly worse performance of 1st graders ($p=0.072$).

[22] In this case, e.g., the effects from the exponentiated coefficients are calculated as follows: (0.999-1+0.643–1)* 100 =-35.8 and (0.645-1)*100=-35.5.

An analysis that counts moves as correct that were made to winning positions even though the treasure was not visible is included in Tables E.6.3 and E.7.3 of the Online Appendices E.6 and E.7. Under this assumption old university students perform significantly better than 6th graders across specifications in part 1. In part 2, the disadvantage female 6th graders have relative to their male classmates turns weakly significant across specifications. Similar to analysis (i), the gender effects in part 1 depend somewhat on the choice of the data set. When using the unhomogenized data set and counting only moves as correct that were made after uncovering the treasure, the disadvantage of female old university students turns significant. Furthermore, when excluding the shortest game $G(m=3, k=4)$, we already find a significantly negative effect of being female for 6th graders (see the Online Appendix E.6).

When pooling observations from part 1 and part 2 in an additional regression we observe similar effects as when considering part 2 alone (see Table E.8.2 in Online Appendix E.8). In this regression we can also include a dummy variable for the observations from part 2. This variable is significant across specifications and indicates an increase of performance of around 20.0 percent due to the repetition supporting Hypothesis 2 ($p<0.001$).

6 Conclusion

To the best of our knowledge, this is the first study which sheds light on the development of the capability to reason backwards in children, adolescents, and young adults. We develop a graphical variant of the race game that is suitable for testing this ability already at an age of 6 years up to an age of 23 years. Behavior is considered in two identical series of race games that allow us to observe improvements of performance. Our findings confirm previous research insofar as, on average, subjects have difficulties conducting backward analyses. But the results presented here reveal that these difficulties diminish with age. In particular, subjects are able to learn how to solve a race game. Differentiating between genders we find significant differences not only regarding the ability to analyze backwards in the first place, but also with respect to learning this ability. While there are no gender differences up to the 4th grade, male 4th graders significantly improve in solving the race game. Accordingly, we find some evidence for performance differences between males and females among 4th graders attending elementary school and among 6th graders attending secondary school. Since there are no differences between males and females among 9th graders (and university students), it seems that females catch up to males after the 6th grade.

Although we took different measures to avoid selection problems (see also footnote 6), the results reported should be taken with some care. As our study is not based on longitudinal data, we cannot exclude the possibility that the results are somehow biased due to selection effects through educational tracking or that the reported effects are cohort effects, not age effects. However, we find some supportive evidence for our observed gender-specific path of development in the data obtained by Harbaugh et al. (2001) who test whether children make rational choices about consumption goods, but do not check for gender differences. Also the results provided by Bishop et al. (2001) who test planning abilities in Tower of Hanoi tasks are in line with our age-related gender effects. In addition, age-related gender effects are found in a meta-study on mathematics performance (Lindberg et al., 2010). The reported gender effects seem to be specifically pronounced in high school, though. Note that our results are not at odds with observations made by Czermak et al. (2011) in static two-person games, who find no relationship between age and the level of strategic sophistication. As they study the behavior of 5^{th}, 7^{th}, 9^{th}, and 11^{th} graders the differences can be attributed to the different age ranges under consideration.

In general, developmental research has offered several explanations for the differences between males and females observed in cognitive tasks. While some interpretations refer to biological causes like innate brain differences or hormonal influences, other explanations are based on environmental causes like socialization practices or stereotypes (for an overview see, e.g., Miller and Halpern, 2014). As both, biological and environmental factors, interact with each other, in recent years an increasing number of studies rely on a broader integrative approach that tries to take into account this interaction. Until now the origin of gender differences in cognitive abilities is still highly debated, however (see, e.g., Guiso et al., 2008).

Biological as well as environmental factors may have caused the age-related gender differences observed in our study. For example, there is some developmental research on planning abilities linking performance in the Tower of London task (in which there appear to be age-related gender effects similar to that observed in our study) to the maturation of the dorsolateral prefrontal cortex (e.g., Kaller et al., 2012). As there is some evidence for gender differences in brain maturation, age-specific gender effects in planning ability (and possibly in the capability to reason backwards) might be related to the differences in maturational trajectories for males and females (see, e.g., the discussion in Unterrainer et al., 2013). Instead or in addition, the age-related different performance of males and females might result from different mathematical ability estimates parents and teachers give to the two genders (e.g., Hyde et al., 2008) and their age-related influence on behavior. Although our games are not

framed as specific math tasks, such estimates might still have affected performance in our study.

Finally, we cannot exclude the possibility that the weight subjects give to winning a game or the familiarity with similar strategic situations vary with age and across gender. As such, our study provides first evidence for the development of the capability to reason backwards. However, more research is needed to isolate the factors that drive this development.

Acknowledgments

We thank William Harbaugh for providing the data from Harbaugh et al. (2001) and Ernan Haruvy and Angela Dorrough for helpful comments. We also benefited from the comments of conference participants in Chicago, Lund, Luxembourg, and Nuremberg. Furthermore, we thank the teachers at the participating schools for their help. Financial support by the Fritz Thyssen Foundation is gratefully acknowledged.

References

Bishop, D. V. M., G. Aamodt-Leeper, C. Creswell, R. McGurk, D. H. Skuse (2001): Individual Differences in Cognitive Planning on the Tower of Hanoi Task: Neuropsychological Maturity or Measurement Error?, *Journal of Child Psychology and Psychiatry* 42 (4), 551–556.

Bornstein, G., T. Kugler, A. Ziegelmeyer (2004): Individual and Group Decisions in the Centipede Game: Are Groups More 'Rational' Players?, *Journal of Experimental Social Psychology* 40(5), 599-605.

Brosig, J., J. P. Reiß (2007): Entry Decisions and Bidding Behavior in Sequential First-Price Procurement Auctions: An Experimental Study, *Games and Economic Behavior* 58(1), 50-74.

Burks, S. V., J. Carpenter, L. Götte, A. Rustichini (2009): Cognitive Skills Affect Economic Preferences, Social Awareness, and Job Attachment, *Proceedings of the National Academy of Science* 106(19), 7745-7750.

Carpenter, J., M. Graham, J. Wolf (2013): Cognitive Ability and Strategic Sophistication, *Games and Economic Behavior* 80, 115-130.

Czermak, S., F. Feri, D. Rützler, M. Sutter (2011): Strategic Sophistication of Adolescents: Evidence from Experimental Normal-Form Games, *IZA Discussion Paper No. 5049*.

Diamond, P., B. Köszegi (2003): Quasi-Hyperbolic Discounting and Retirement, *Journal of Public Economics* 87, 1839-1872.

Dufwenberg, M., R. Sundaram, D. J. Butler (2010): Epiphany in the Game of 21, *Journal of Economic Behavior & Organization* 75 (2), 132-143.

Fey, M., R. D. McKelvey, T. R. Palfrey (1996): An Experimental Study of Constant-Sum Centipede Games, *International Journal of Game Theory* 25(3), 269-287.

Fischbacher, U. (2007): z-Tree: Zurich Toolbox for Ready-made Economic Experiments, *Experimental Economics* 10(2), 171-178.

Gneezy, U., A. Rustichini, A. Vostroknutov (2010): Experience and Insight in the Race Game, *Journal of Economic Behavior & Organization* 75(2), 144-155.

Greiner, B. (2004): An Online Recruitment System for Economic Experiments, in: K. Kremer, V. Macho (eds.): Forschung und wissenschaftliches Rechnen 2003. *GWDG Bericht* 63, 79-93.

Griffin, J. (2005): *Intcens: Stata Module to Perform Interval-censored Survival Analysis*. URL: http://ideas.repec.org/c/boc/bocode/s453501.html.

Guiso, L., F. Monte, P. Sapienza, L. Zingales (2008). Culture, Gender, and Math, *Science* 320(5880), 1164-1165

Harbaugh, W. T., K. Krause, T. R. Berry (2001): GARP for Kids: on the Development of Rational Choice Behavior, *American Economic Review* 91(5), 1539-1545.

Hawes, D. R., A. Vostroknutov, A. Rustichini (2012): Experience and Abstract Reasoning in Learning Backward Induction, *Frontiers in Neuroscience* 6 (23), 1-13.

Hyde, J. S., S. M. Lindberg, M. C. Linn, A. B. Ellis, C. C. Williams (2008): Gender Similarities Characterize Math Performance, *Science* 321(5888), 494-495.

Johnson, E. J., C. Camerer, S. Sen, T. Rymon (2002): Detecting Failures of Backward Induction: Monitoring Information Search in Sequential Bargaining, *Journal of Economic Theory* 104(1), 16-47.

Kaller, C. P., K. Heinze, I. Mader, J. M. Unterrainer, B. Rahm, C. Weiller, L. Köstering (2012): Linking Planning Performance and Gray Matter Density in Mid-Dorsolateral Prefrontal Cortex: Moderating Effects of Age and Sex, *Neuroimage* 63, 1454-1463.

Klein, J. P., M. L. Moeschberger (2003): *Survival Analysis - Techniques for Censored and Truncated Data*. Springer.

Laibson, D. (1997): Golden Eggs and Hyperbolic Discounting, *Quarterly Journal of Economics* 112(2), 443-478.

Levitt, S. D., J. A. List, S. E. Sadoff (2011): Checkmate: Exploring Backward Induction among Chess Players, *American Economic Review* 101(2), 975-90.

Lindberg, S. M., J. S. Hyde, J. L. Petersen, M. C. Linn (2010): New Trends in Gender and Mathematics Performance: A Meta-Analysis, *Psychological Bulletin* 136(6), 1123-1135.

LWL (2009): Rundschreiben Nr. 27/2009, LWL-Landesjugendamt, http://www.lwl.org.

McCormack, T., C. M. Atance (2011): Planning in Young Children: A Review and Synthesis, *Developmental Review* 31, 1-31.

McKelvey, R. D., T. R. Palfrey (1992): An Experimental Study of the Centipede Game, *Econometrica* 60(4), 803-836.

McKinney, C. N., J. B. Van Huyck (2006): Does Seeing More Deeply into a Game Increase One's Chances of Winning?, *Experimental Economics* 9(3), 297-303.

Miller, D. I., D. F. Halpern (2014): The new Science of Cognitive Sex Differences, *Trends in Cognitive Sciences* 18(1), 37-45.

Nagel, R., F. F. Tang (1998): Experimental Results on the Centipede Game in Normal Form: An Investigation on Learning, *Journal of Mathematical Psychology* 42(2-3), 356-384.

Parco, J. E., A. Rapoport, W. E. Stein (2002): Effects of Financial Incentives on the Breakdown of Mutual Trust, *Psychological Science* 13(3), 292-297.

Rapoport, A., W. E. Stein, J. E. Parco, T. E. Nicholas (2003): Equilibrium Play and Adaptive Learning in a Three-Person Centipede Game, *Games and Economic Behavior* 43(2), 239-265.

Royston, P., M. K. B. Parmar (2002): Flexible Parametric Proportional-hazards and Proportional-odds Models for Censored Survival Data, with Application to Prognostic Modelling and Estimation of Treatment Effects, *Statistics in Medicine* 21, 2175–2197

Schneider, W., J. Stefanek, F. Niklas (2008). Development of Intelligence and Thinking, in: W. Schneider, M. Bullock (eds.): *Human Development from Early Childhood to Early Adulthood: Findings from a 20 Year Longitudinal Study*, Lawrence Erlbaum Assoc Inc, pp. 7-33.

Stern, E. (1999). Development of Mathematical Competencies, in: F.E. Weinert, W. Schneider (eds.): *Individual development from 3–12: Findings from the Munich Longitudinal Study,* Cambridge University Press, pp. 154–170.

Stern, E. (2008). The Development of Mathematical Competencies: Sources of Individual Differences and their Developmental Trajectories, in: W. Schneider, M. Bullock (eds.): *Human Development from Early Childhood to Early Adulthood: Findings from a 20 Year Longitudinal Study*, Lawrence Erlbaum Assoc Inc, pp. 221-236.

Unterrainer, J. M., C. P. Kaller, S. V. Loosli, K. Heinze, N. Ruh, M. Paschke-Müller, R. Rauh, M. Biscaldi, B. Rahm (2014): Looking Ahead from age 6 to 13: A Deeper Insight into the Development of Planning Ability. Forthcoming in: *British Journal of Psychology*.

Unterrainer, J. M., N. Ruh, S. V. Loosli, K. Heinze, B. Rahm, C. P. Kaller (2013): Planning Steps Forward in Development: In Girls Earlier than in Boys, *PLoS ONE* 8(11), e80772.

Appendix A: Description of variables

Variable	Description
ID	Unique subject number
Age (months)	Age in months
Grade 1	Is subject from grade 1 (restricted sample), 1 = yes, 0 = no
Grade 4	Is subject from grade 4 (restricted sample), 1 = yes, 0 = no
Grade 6	Is subject from grade 6 (restricted sample), 1 = yes, 0 = no
Grade 9	Is subject from grade 9 (restricted sample), 1 = yes, 0 = no
Student (young)	Is subject a "young" university student (restricted sample), 1 = yes, 0 = no
Student (old)	Is subject an "old" university student (restricted sample), 1 = yes, 0 = no
Grade (12)	Grade when the sample is restricted to a maximum age range of 12 months at every level, 1 = grade 1, 4 = grade 4, 6 = grade 6, 9 = grade 9, 14 = young university student, 15 = old university student
Grade (full)	Grade without sample restriction, 1 = grade 1, 4 = grade 4, 6 = grade 6, 9 = grade 9, 14 = young university student, 15 = old university student
Female	Gender, 1 = female, 0 = male
Uncover treasure	Did the subject uncover the treasure before the very first move in game 1? 1 = yes, 0 = no
Mark German	School mark in German on a scale from 1 = very poor to 15 = very good (for university students the last school mark is used)
Mark math	School mark in math on a scale from 1 = very poor to 15 = very good (for university students the last school mark is used)
Fairness	Reply to the question 'Do you think most people…' on a scale from 0 = '… would exploit you if they were given the chance' to 1 = '… would try to treat you in a fair way'.
Part 2	Is the analyzed move in part 2 of the experiment, 1 = yes, 0 = no
Patience	Reply to the question 'How do you see yourself: Are you normally an impatient person or do you usually exercise patience?' on a scale from 0 = 'very impatient' to 10 = 'very patient'.
Risk	Reply to the question 'How do you see yourself: Do you normally take risks or try to avoid them?' on a scale from 0 = 'do not take risks at all' to 10 = 'take a lot of risks'.
Trust	Reply to the question 'What is your opinion on the following statement: In general, one can trust people' on a scale from 1 = 'fully disagree' to 4 = 'fully agree'.
Wins (1)	Number of games a subject wins in part 1
Wins (2)	Number of games a subject wins in part 2

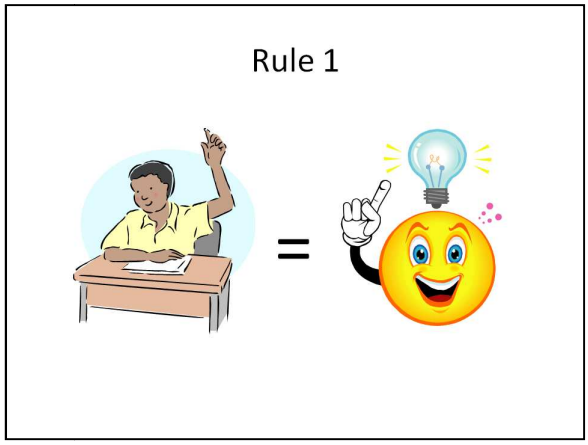
Table A: Description of variables

FOR ONLINE PUBLICATION

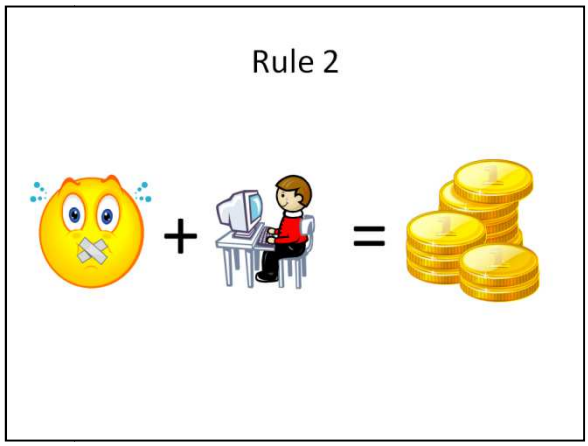
Online Appendix B: Instructions (translated from German)

Hello and welcome! Today you are taking part in an experiment in which you will be able to earn money. How much money you will earn depends on your decisions.

Important: All your decisions are made anonymously. Nobody will be able to link the choices you made to your name. We will tell you in a moment what the experiment is about. First of all there are two important rules:



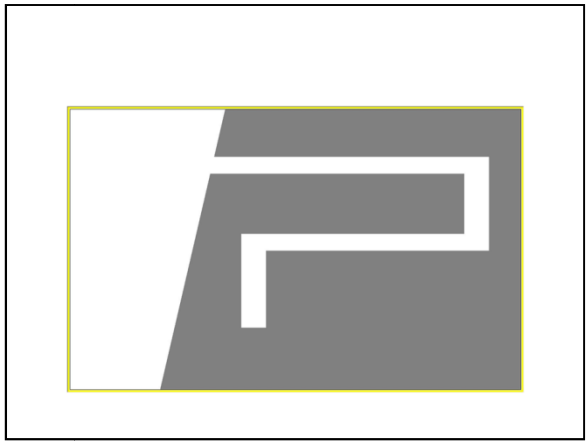
1. Signal us, if you do not understand something. We want you to have a perfect understanding of everything!



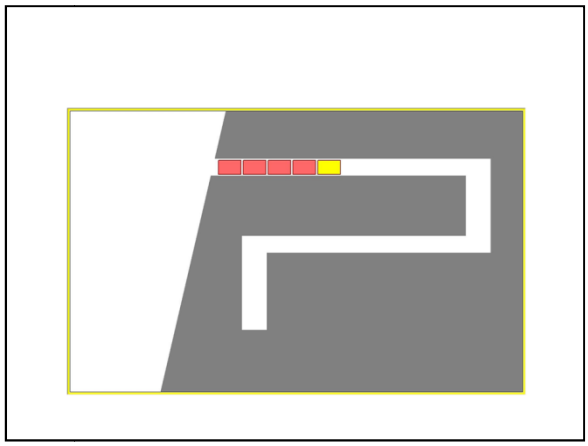
2. It is not allowed to talk to other participants. If you, however, talk to another participant you will be immediately excluded from this experiment. Consequently you will also earn no money in this case.

Now let's return to the rules of the game. Several times each of you will individually play a game against the computer. The more often you win against the computer, the more money you will earn. How does the game look like?

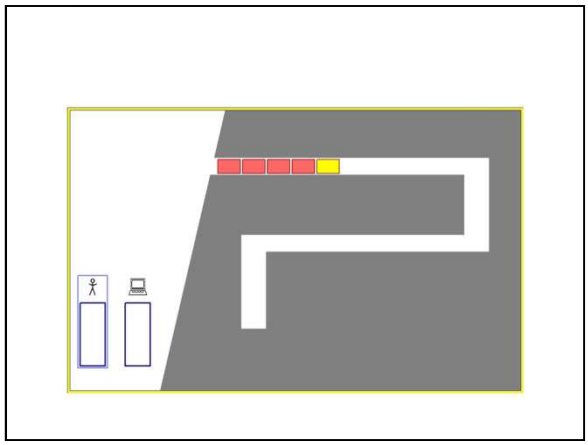
The computer challenges you. The goal of the game is to reach and collect a treasure.



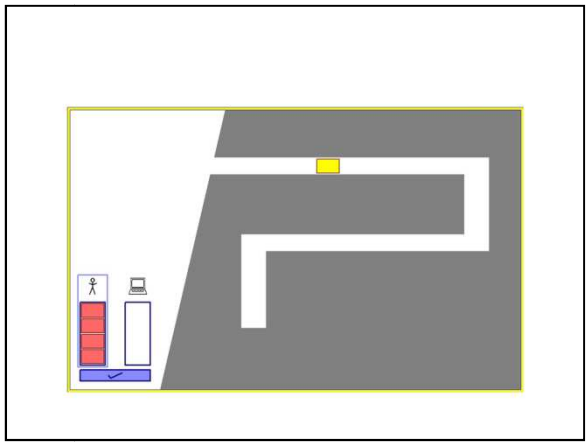
The treasure, which looks like a yellow square, has been buried by the computer in a cave, which has only one entry. You and the computer can reach the treasure only by using the entrance of the cave.



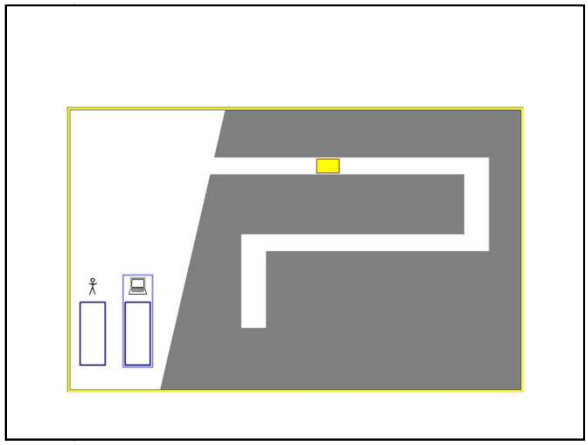
Unfortunately, the computer has blocked the passage from the entry to the treasure with one or more red stones. (This means that behind the treasure there are no more red stones but only air.) To reach the treasure you have to remove the red stones. The stones can only be removed by carrying them out through the entrance of the cave. This means that you and the computer can only move the stone, which is the closest to the entrance.



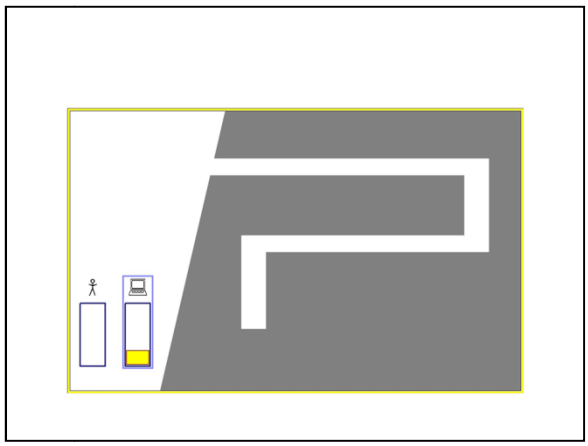
You can remove the stones by packing them into your box. This box (the blue rectangle on the left side) only fits one, two, three, or at most four stones.



By removing the stones you alternate with your rival, the computer.



The computer can also remove stones and also has a box (the blue rectangle on the right side) that fits four stones at most.

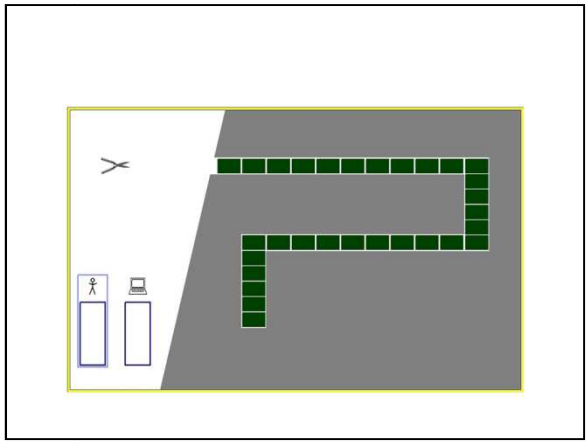


After every move the stones in your box and the computer's box disappear. The winner is the player who packs the treasure in his box first. In this game you are always the first who can pack stones into your box and remove them. Thereafter, it is the computer's turn. Same as you, the computer tries to win the game and to put the treasure into his box.

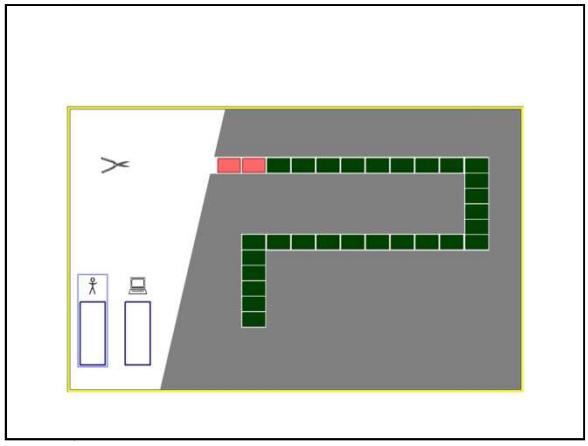
You can pick a stone by clicking on it with the mouse, holding the button and pulling the stone into your box. (If you have put more stones in your box than you wanted to, you can pull the stones back into the cave.) When there are as many stones in your box as you want to remove, you have to click on the blue checkmark button and the stones disappear. Then it is the computer's turn and you can observe how many stones the computer removes.

[Video]

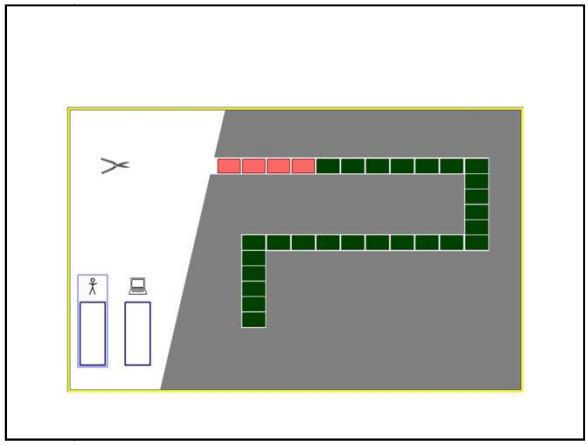
You take turns with the computer until one of you removes the treasure and wins. You can only win by packing the stone into your box and removing it. Same as you, the computer must remove at least one stone at each turn.



There is a special feature: After the computer has hidden the treasure it has blocked your view to the cave with bushes, that look like green squares. The computer knows what is hidden behind each green bush. If you want to see what is behind the green bushes as well, you just have to click on the hedge trimmer.



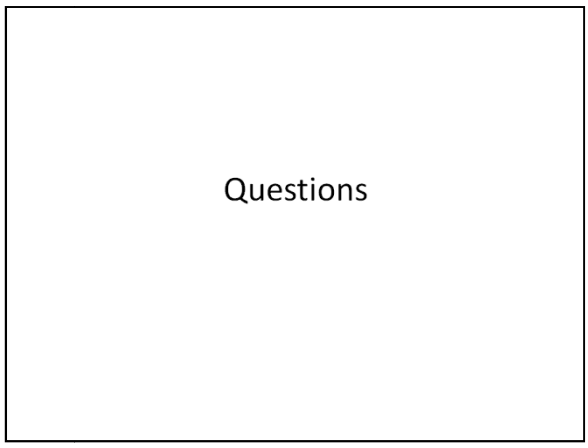
Starting from the entry of the cave you can remove two adjacent bushes by clicking on the hedge trimmer once.



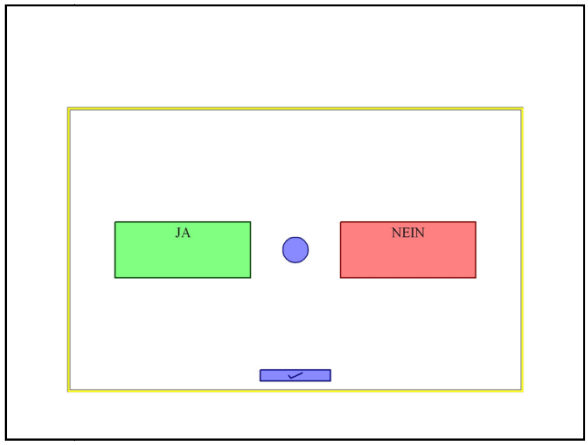
In each move you can remove as many green bushes as you want. Behind every green bush there can be either a red stone, or nothing, or the treasure. As mentioned before, you play several games consecutively. These games differ from each other only in the number of red stones, blocking your way to the treasure. For each game you win you will receive 5 [amount depending on age group] Euro.

When you entered this room you received a card with your code name. Please keep it safe. At the end of the experiment you have to enter your code name in the computer. You also need your card to collect your payoff.

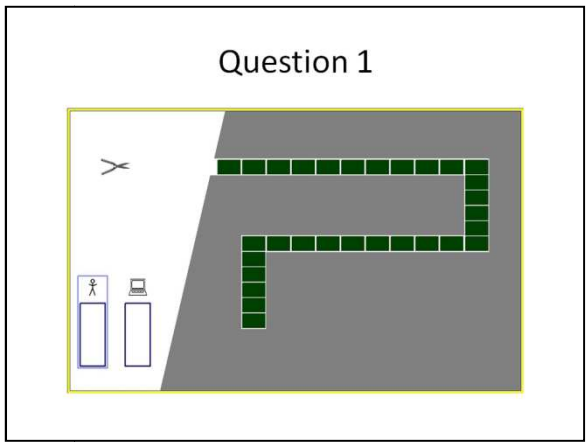
At the end of the experiment your respective payoffs will be calculated. After this the cash desk in the corridor outside the laboratory opens. There you can collect a closed envelope containing your payoff by showing your card. The cashier does not know what is inside these envelopes. Please collect your payoffs immediately after the experiment. [Payoff description of the instructions for university students.]



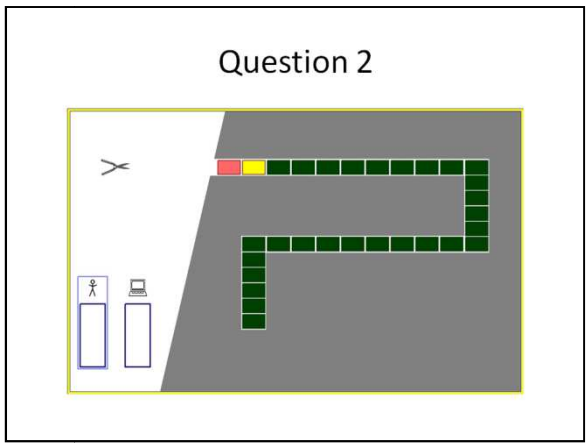
Before we start we will ask some questions so that we can help you better to understand the game.



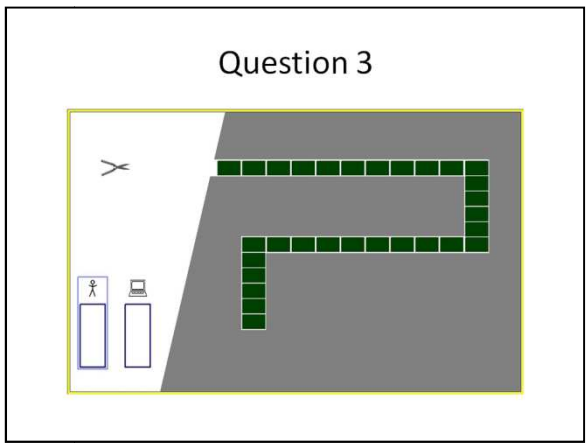
Please answer the questions with “Yes” or “No” by pulling the blue ball, which will appear in front of you on the monitor, into the green or the red area.



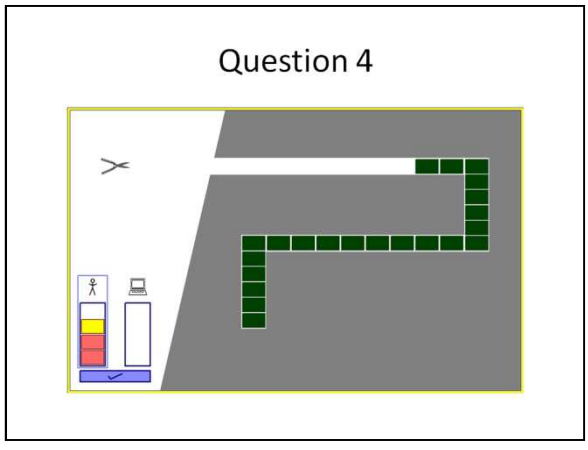
1) Please have a look at the following game. Does the computer know behind which green bush the treasure lies?



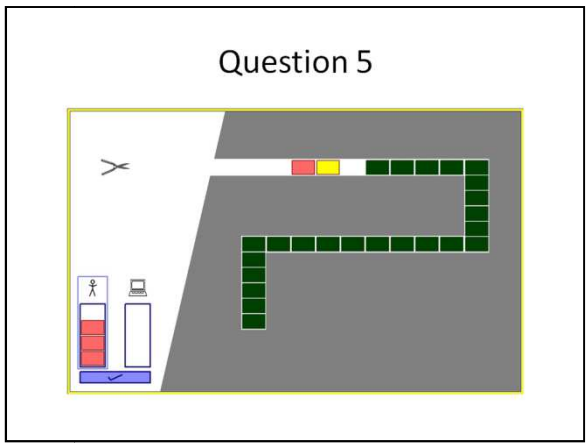
2) Please have a look at the following game. Do you see where the treasure is?



3) Please have a look at the following game. Are you allowed to remove all green bushes with the hedge trimmer now?



4) Please have a look at the following game. Is it correct, that the computer is winning the game?



5) Please have a look at the following game. You want to pack two stones into your box in order to win. Is this possible?

FOR ONLINE PUBLICATION

Online Appendix C: Ex-post questionnaire (translated from German)

We would be very pleased if you could take a few minutes to answer some questions. You can start the questionnaire by clicking "continue". As soon as every participant completed the questionnaire, participants are paid off one by one.

1. How did you make your decisions? Please explain briefly.

2. How do you see yourself: Do you normally take risks or try to avoid them?

   Do not take risks at all ☐☐☐☐☐☐☐☐☐☐☐ Take a lot of risks

3. How do you see yourself: Are you normally an impatient person or do you usually exercise patience?

   Very impatient ☐☐☐☐☐☐☐☐☐☐☐ Very patient

4. What is your opinion on the following statement: In general, one can trust people.

   ☐ Fully agree
   ☐ Rather agree
   ☐ Rather disagree
   ☐ Fully disagree

5. Do you think most people...

   ☐ … would exploit you if they were given the chance?
   ☐ … would try to treat you in a fair way?

6. Please think back to your school days: Which mark did you reach in math on your last school report? What kind of course did you take part in?

   Math

   ☐ 1 – very good (13 – 15 points)
   ☐ 2 – good (10 – 12 points)
   ☐ 3 – satisfactory (7 – 9 points)
   ☐ 4 – sufficient (4 – 6 points)
   ☐ 5 – unsatisfactory (1 – 3 points)
   ☐ 6 – very poor (0 points)

   Course

   ☐ Basic course
   ☐ Advanced course
   ☐ Other course
   ☐ Do not know

   If you indicated "other course": How was this course named?

7. Please think back to your school days: Which mark did you reach in German on your last school report? What kind of course did you take part in?

   German

   - ☐ 1 – very good (13 – 15 points)
   - ☐ 2 – good (10 – 12 points)
   - ☐ 3 – satisfactory (7 – 9 points)
   - ☐ 4 – sufficient (4 – 6 points)
   - ☐ 5 – unsatisfactory (1 – 3 points)
   - ☐ 6 – very poor (0 points)

   Course

   - ☐ Basic course
   - ☐ Advanced course
   - ☐ Other course
   - ☐ Do not know

   If you indicated “other course”: How was this course named?

8. Please specify your gender:

   - ☐ Male
   - ☐ Female

9. Please indicate your year and month of birth:

   Year:
   Month (1 – 12):

10. Which is your field of study (field and degree)?

11. In which semester are you in this field of study?

This was the last page of the questionnaire. Thank you very much!

FOR ONLINE PUBLICATION

Online Appendix D: Data

ID	*Age (months)*	*Grade (12)*	*Grade (full)*	*Female*	*Wins (I)*	*Wins (II)*	*Mark math*	*Mark German*	*Risk*	*Patience*	*Trust*	*Fair*	*Uncover (1st game)*
1	247	14	14	1	2	3	10	13	7	2	2	0	1
2	308		15	1	3	4	13	13	1	8	2	0	1
3	240	14	14	1	2	3	13	12	7	6	3	0	0
4	262		14	0	3	3	6	12	8	9	2	0	1
5	250	14	14	0	2	2	10	13	6	5	3	1	1
6	293		15	1	1	2	12	12	2	4	2	0	0
7	249	14	14	1	3	3	8	12	1	2	2	0	1
8	290		15	1	2	3	8	14	2	9	2	0	0
9	313		15	1	3	3	8	10	8	5	3	1	1
10	249	14	14	0	3	3	9	6	7	2	2	1	1
11	257		14	1	2	5	9	13	7	1	3	1	1
12	295		15	0	3	3	15	8	8	3	3	0	1
13	311		15	0	3	3	7	12	8	9	4	1	0
14	312		15	0	2	3	10	12	6	9	3	1	0
15	246	14	14	1	3	3	12	13	4	3	3	0	1
16	269		14	0	4	3	12	12	1	3	3	1	1
17	263		14	1	3	3	11	14	7	10	4	1	0
18	274		15	0	3	3	13	12	2	7	2	0	0
19	253		14	0	3	4	12	10	4	2	3	1	1
20	261		14	0	2	3	13	8	2	2	2	0	1
21	241	14	14	1	2	4	10	12	7	4	3	0	1
22	240	14	14	0	1	2	11	10	6	1	3	1	1
23	272		15	0	3	2	10	10	7	4	3	1	0
24	243	14	14	1	3	3	12	12	7	3	3	0	0
25	127	4	4	1	2	1	11	11					1
26	124	4	4	1	3	3	11	11					1
27	123	4	4	1	1	1	8	8					0
28	126	4	4	1	2	3	11	14					1
29	121	4	4	1	3	2	8	11					1
30	130	4	4	1	1	1	5	8					1
31	126	4	4	1	2	1	5	11					0
32	126	4	4	1	1	1	11	11					1
33	121	4	4	0	2	2	11	11					1
34	126	4	4	0	1	2	11	11					0
35	123	4	4	0	2	2	8	11					0
36	126	4	4	0	1	1	11	11					1
37	118		4	0	2	3	11	11					0
38	121	4	4	0	4	4	14	11					1
39	121	4	4	0	2	3	8	5					0
40	129	4	4	1	1	1	8	11					1
41	126	4	4	1	3	2	8	11					1
42	140		4	1	1	1	8	11					1
43	130	4	4	1	1	1	5	11					1
44	128	4	4	1	1	2	11	14					0
45	120	4	4	1	1	2	11	11					1
46	119	4	4	1	2	1	8	8					1
47	128	4	4	0	1	2	11	11					0
48	121	4	4	0	2	3	8	11					1
49	127	4	4	0	3	3	11	8					1
50	129	4	4	0	1	2	5	8					1
51	128	4	4	0	3	3	11	11					1
52	127	4	4	0	2	3	11	11					1
53	129	4	4	0	3	3	11	14					1
54	120	4	4	0	2	3	11	11					0
55	92	1	1	1	2	3							1
56	83	1	1	1	2	1							0
57	85	1	1	1	1	2							0
58	82	1	1	1	1	2							1
59	91	1	1	1	2	3							1
60	93	1	1	1	2	1							1
61	91	1	1	1	2	2							1
62	90	1	1	1	1	1							1
63	92	1	1	0	1	2							1
64	84	1	1	0	1	1							1
65	88	1	1	0	2	1							1
66	86	1	1	0	2	2							1
67	82	1	1	1	1	3							0
68	95		1	1	1	1							0
69	82	1	1	1	2	1							1

ID	*Age (months)*	*Grade (12)*	*Grade (full)*	*Female*	*Wins (I)*	*Wins (II)*	*Mark math*	*Mark German*	*Risk*	*Patience*	*Trust*	*Fair*	*Uncover (1^{st} game)*
70	83	1	1	1	1	1							1
71	82	1	1	1	1	2							1
72	82	1	1	1	1	2							0
73	93	1	1	1	1	2							1
74	90	1	1	1	1	1							1
75	89	1	1	1	1	1							1
76	89	1	1	1	1	2							0
77	93	1	1	0	1	1							0
78	88	1	1	1	1	0							0
79	84	1	1	0	1	1							1
80	82	1	1	0	1	3							1
81	90	1	1	0	2	2							1
82	93	1	1	0	1	3							1
83	91	1	1	0	1	2							1
84	84	1	1	1	0	1							0
85	186	9	9	0	2	2	10	8	8	1	3	1	0
86	194		9	1	1	2	10	12	6	2	3	1	1
87	189	9	9	1	2	4	8	12	5	4	3	0	0
88	183	9	9	0	3	0	14	12	3	9	3	1	1
89	190	9	9	0	3	3	6	10	5	7	4	1	1
90	184	9	9	1	1	1	10	8	5	5	3	1	0
91	190	9	9	0	2	3	12	10	3	4	4	1	0
92	185	9	9	1	3	3	14	10	7	1	4	1	0
93	186	9	9	0	0	1	10	10	3	7	3	1	0
94	188	9	9	0	2	3	8	10	7	3	2	1	0
95	191		9	1	0	3	10	8	5	5	2	0	0
96	186	9	9	0	2	3	6	12	3	6	4	1	1
97	196		9	1	2	3	12	10	4	8	2	1	1
98	203		9	0	4	3	12	8	2	9	3	1	1
99	188	9	9	0	2	2	10	6	8	7	2	1	1
100	188	9	9	1	0	2	4	6	5	6	4	1	0
101	184	9	9	1	2	2	8	8	7	4	3	1	1
102	142	6	6	1	2	3	11	11	2	8	3	1	1
103	157		6	1	1	1	5	8	4	9	3	1	0
104	148	6	6	0	2	1	5	5	3	6	3	1	1
105	152	6	6	0	2	2	8	11	4	0	3	1	1
106	150	6	6	1	1	2	8	11	4	8	4	1	1
107	151	6	6	0	3	4	2	5	4	8	4	1	0
108	153	6	6	1	1	2	8		3	8	3	1	0
109	148	6	6	0	2	3	8	8	3	8	3	1	1
110	147	6	6	1	1	1	5	8	3	7	3	1	0
111	159		6	1	0	1	5	8	8	3	3	1	0
112	153	6	6	1	2	1	11	11	5	7	2	1	0
113	145	6	6	1	2	1	8	11	6	9	3	1	1
114	144	6	6	1	3	2	11	11	4	7	3	1	1
115	152	6	6	1	2	3	14	11	1	5	3	1	0
116	164		6	1	1	2	2	5	6	7	3	1	1
117	187	9	9	0	1	1	4	10	6	5	3	0	0
118	180	9	9	1	1	2	4	6	1	5	4	0	1
119	182	9	9	0	3	4	12	10	3	6	3	1	0
120	184	9	9	1	1	3	10	12	6	9	2	1	0
121	179	9	9	0	2	3	10	6	6	6	3	0	0
122	189	9	9	0	2	3	10	12	6	4	3	1	1
123	181	9	9	0	0	2	10	10	3	6	2	0	0
124	190	9	9	0	3	3	8	10	5	2	3	1	1
125	194		9	1	1	2	4	4	5	4	3	0	0
126	193		9	0	0	1	6	4	8	5	3	0	0
127	180	9	9	0	3	2	8	12	8	2	2	1	1
128	190	9	9	1	1	2	12	12	7	9	3	1	0
129	185	9	9	1	2	4	10	12	9	1	3	1	1
130	152	6	6	1	2	2	11	11	5	8	3	1	0
131	143	6	6	0	4	4	14	11	7	9	3	1	1
132	148	6	6	1	2	1	11	11	5	8	3	0	0
133	157		6	0	4	2	11	8	6	5	4	1	1
134	144	6	6	1	2	1	2	8	6	4	3	0	1
135	151	6	6	0	2	2	11	11	6	5	3	1	1
136	143	6	6	0	2	2	11	14	7	5	3	1	0
137	145	6	6	0	1	1	11	14	3	3	2	1	0
138	152	6	6	0	0	3	8	11	7	5	3	1	0
139	155		6	0	2	3	14	11	8	1	2	0	1
140	149	6	6	0	2	3	11	11	8	2	3	1	1
141	145	6	6	0	2	3	8	11	2	9	3	1	1
142	148	6	6	0	3	3	11	8	6	6	3	1	1
143	144	6	6	0	3	2	14	11	1	3	3	0	1
144	144	6	6	0	2	4	11	11	4	5	3	1	0

ID	*Age (months)*	*Grade (12)*	*Grade (full)*	*Female*	*Wins (I)*	*Wins (II)*	*Mark math*	*Mark German*	*Risk*	*Patience*	*Trust*	*Fair*	*Uncover (1^{st} game)*
145	242	14	14	1	3	2	10	13	7	2	2	1	1
146	246	14	14	0	3	3	8	10	4	5	3	0	1
147	246	14	14	1	1	2	11	12	3	5	2	1	1
148	243	14	14	0	3	3	12	12	7	2	2	1	1
149	240	14	14	1	2	2	10	13	3	9	2	1	1
150	242	14	14	1	1	2	12	9	2	8	2	1	0
151	240	14	14	0	3	4	14	15	7	5	3	1	1
152	244	14	14	0	3	3	8	13	8	2	2	1	1
153	245	14	14	0	1	1	10	12	5	2	3	1	0
154	241	14	14	1	2	1	10	10	8	3	2	0	0
155	248	14	14	0	3	4	11	8	7	2	3	0	1
156	242	14	14	1	3	2	10	11	4	2	3	1	0
157	278	15	15	1	1	1	6	15	3	6	2	0	1
158	277	15	15	1	2	3	10	9	9	9	3	1	1
159	286	15	15	0	2	3	10	12	7	4	2	0	1
160	283	15	15	1	1	1	10	13	2	0	2	0	0
161	287	15	15	0	3	6	9	8	8	4	2	0	1
162	276	15	15	1	3	4	12	11	2	2	3	1	1
163	281	15	15	0	3	3	8	10	2	1	2	0	1
164	283	15	15	1	3	3	12	15	1	3	3	1	1
165	284	15	15	1	2	5	10	13	2	5	3	1	1
166	280	15	15	1	1	3	6	13	5	2	3	0	1
167	277	15	15	1	2	3	12	13	3	9	3	0	1
168	277	15	15	0	3	5	10	13	3	2	3	0	1
169	276	15	15	1	2	5	15	12	2	1	3	1	1
170	277	15	15	0	1	1	11	8	8	2	3	1	1
171	287	15	15	0	3	3	10	14	5	2	2	1	1
172	276	15	15	0	4	3	10	12	3	4	3	1	0
173	281	15	15	0	3	4	10	10	8	7	4	1	1
174	287	15	15	0	2	1	11	12	7	3	3	1	0
175	282	15	15	1	4	3	12	12	2	8	2	0	0

Table D: Data

FOR ONLINE PUBLICATION

Online Appendix E: Additional regressions

The following subsections present additional regression results for the dependent variables analyzed in the paper. For the regressions presented in detail in the main part of the paper this appendix also includes robustness checks that are based on different data sets. In this study we use the following data sets:

(i) the restricted data set with subjects falling into 12-month ranges (used in the main part),
(ii) the restricted data set without game $G(m=3, k=4)$,
(iii) the unrestricted data set and
(iv) the unrestricted data set without game $G(m=3, k=4)$.

These data sets yield the summary statistics presented in Table E.1. The following analyses indicate the respective data set the regression is run for.

	N	Female	Minimum Age	Maximum Age
Grade 1				
(i) and (ii)	29	66%	06 y 10 m	07 y 09 m
(iii) and (iv)	30	67%	06 y 10 m	07 y 11 m
Grade 4				
(i) and (ii)	28	50%	09 y 11 m	10 y 10 m
(iii) and (iv)	30	50%	09 y 10 m	11 y 08 m
Grade 6				
(i) and (ii)	25	44%	11 y 10 m	12 y 09 m
(iii) and (iv)	30	47%	11 y 10 m	13 y 08 m
Grade 9				
(i) and (ii)	24	38%	14 y 11 m	15 y 10 m
(iii) and (iv)	30	43%	14 y 11 m	16 y 11 m
University young				
(i) and (ii)	21	57%	20 y 00 m	20 y 10 m
(iii) and (iv)	27	52%	20 y 00 m	22 y 05 m
University old				
(i) and (ii)	19	53%	23 y 00 m	23 y 11 m
(iii) and (iv)	28	50%	22 y 08 m	26 y 01 m

Table E.1: Data sets (i), (ii), (iii) and (iv)

Online Appendix E.1: Improvement between parts within games

The results presented in section 5.2 of the paper suggest that performance generally increases from part 1 to part 2. This effect is significant for 1st graders, 9th graders and old university students. In addition we ran the following ordered probit regressions to study how the improvement differs across age groups in the four games that exhibit the largest variation in performance, i.e. $G(m=8, k=4)$, $G(m=11, k=4)$, $G(m=19, k=4)$, and $G(m=21, k=4)$.

The dependent variable takes the values -1, 0, and 1 indicating whether subjects won a specific game in part 1 but lost in part 2, won or lost the game in both parts, or won the game in part 2 but lost in part 1. The results in Tables E.1.1 based on data set (i) reveal significant age differences for games $G(m=19, k=4)$ and $G(m=21, k=4)$. They suggest that the increase in performance for game $G(m=19, k=4)$ is particularly driven by 1st graders while the increase in performance for game $G(m=21, k=4)$ is particularly driven by young and old university students.

	Ordered Probit - dependent variable: improvement between parts							
	$G(m=19, k=4)$		$G(m=8, k=4)$		$G(m=11, k=4)$		$G(m=21, k=4)$	
Grade 1	0.755**	(0.353)	0.188	(0.335)	-0.301	(0.326)	0.318	(0.525)
Grade 4	0.200	(0.352)	0.281	(0.339)	-0.296	(0.328)	-0.548	(0.469)
Grade 9	0.247	(0.371)	0.209	(0.354)	0.318	(0.342)	0.318	(0.508)
Uni young	-0.427	(0.386)	0.340	(0.365)	-0.111	(0.351)	1.184**	(0.560)
Uni old	0.000	(0.391)	0.103	(0.380)	0.187	(0.364)	1.756***	(0.551)
Adj. R-squared	0.056		0.006		0.023		0.191	
N	146		146		146		146	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.1.1: Regressions on improvement across games based on data set (i)

Online Appendix E.2: Number of games won in part 1

In order to control for subject-specific differences that are not captured by grade or gender variables, we also present OLS-regressions in the paper that include the number of games won as the dependent variables and additional independent variables that control for uncovering behavior, school grades, and self-reported personality attitudes.

The regressions of analysis (i) in the paper are based on data set (i). Below we present the OLS-regressions for the remaining three data sets with the number of games won in part 1 as the dependent variable in Tables E.2.1a, E.2.1b, and E.2.1c. As an additional robustness check we also present an ordered probit regression based on data set (i) in Table E.2.2.

	OLS - dependent variable: number of games won in part 1					
	(1)		(2)		(3)	
Grade 1	-1.005***	(0.308)				
Grade 4	-0.143	(0.279)	-0.200	(0.292)		
Grade 9	-0.019	(0.275)	0.014	(0.288)	0.003	(0.294)
Uni young	0.143	(0.317)	0.064	(0.333)	0.088	(0.359)
Uni old	0.405	(0.316)	0.371	(0.331)	0.370	(0.356)
Grade 1 x female	0.163	(0.291)				
Grade 4 x female	-0.336	(0.280)	-0.202	(0.297)		
Grade 6 x female	-0.239	(0.299)	-0.228	(0.32)	-0.241	(0.333)
Grade 9 x female	-0.420	(0.312)	-0.388	(0.326)	-0.421	(0.335)
Uni young x female	-0.087	(0.328)	-0.096	(0.345)	-0.110	(0.365)
Uni old x female	-0.575*	(0.339)	-0.617*	(0.359)	-0.620	(0.382)
Uncover treasure	0.353**	(0.136)	0.399**	(0.158)	0.385**	(0.185)
Mark German			-0.009	(0.039)	-0.013	(0.045)
Mark math			0.091***	(0.032)	0.090**	(0.037)
Fairness					-0.103	(0.205)
Patience					-0.001	(0.037)
Risk					-0.001	(0.042)
Trust					0.096	(0.159)
Constant	0.988***	(0.216)	0.183	(0.456)	0.067	(0.807)
Adj. R-squared	0.258		0.212		0.212	
N	146		116		88	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.2.1a: Regressions part 1 based on data set (ii)

	OLS - dependent variable: number of games won in part 1					
	(1)		(2)		(3)	
Grade 1	-1.062***	(0.327)				
Grade 4	-0.137	(0.291)	-0.168	(0.299)		
Grade 9	-0.136	(0.283)	-0.064	(0.292)	-0.041	(0.295)
Uni young	0.242	(0.303)	0.179	(0.313)	0.344	(0.325)
Uni old	0.525*	(0.296)	0.485	(0.306)	0.577*	(0.317)
Grade 1 x female	0.107	(0.315)				
Grade 4 x female	-0.505*	(0.296)	-0.361	(0.309)		
Grade 6 x female	-0.543*	(0.298)	-0.335	(0.316)	-0.394	(0.325)
Grade 9 x female	-0.647**	(0.298)	-0.621**	(0.306)	-0.613*	(0.309)
Uni young x female	-0.145	(0.315)	-0.155	(0.327)	-0.184	(0.334)
Uni old x female	-0.646**	(0.306)	-0.670**	(0.319)	-0.694**	(0.339)
Uncover treasure	0.525***	(0.134)	0.543***	(0.150)	0.567***	(0.170)
Mark German			0.005	(0.036)	0.012	(0.040)
Mark math			0.093***	(0.029)	0.096***	(0.032)
Fairness					0.003	(0.189)
Patience					0.035	(0.032)
Risk					-0.016	(0.038)
Trust					0.275*	(0.145)
Constant	1.889***	0.222	0.906**	(0.435)	-0.137	(0.723)
Adj. R-squared	0.322		0.313		0.361	
N	175		144		114	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.2.1b: Regressions part 1 based on data set (iii)

	OLS - dependent variable: number of games won in part 1					
	(1)		(2)		(3)	
Grade 1	-1.092	(0.301)				
Grade 4	-0.213	(0.267)	-0.240	(0.276)		
Grade 9	-0.055	(0.260)	0.006	(0.269)	0.015	(0.267)
Uni young	0.215	(0.279)	0.163	(0.289)	0.321	(0.294)
Uni old	0.445	(0.272)	0.410	(0.282)	0.497*	(0.287)
Grade 1 x female	0.162	(0.290)				
Grade 4 x female	-0.408	(0.272)	-0.276	(0.285)		
Grade 6 x female	-0.430	(0.274)	-0.299	(0.291)	-0.381	(0.294)
Grade 9 x female	-0.606**	(0.274)	-0.584**	(0.282)	-0.574**	(0.280)
Uni young x female	-0.198	(0.289)	-0.207	(0.302)	-0.235	(0.302)
Uni old x female	-0.625**	(0.281)	-0.641**	(0.294)	-0.654**	(0.307)
Uncover treasure	0.373***	(0.123)	0.381***	(0.138)	0.400**	(0.154)
Mark German			0.002	(0.033)	0.001	(0.036)
Mark math			0.084***	(0.027)	0.083***	(0.029)
Fairness					0.006	(0.171)
Patience					0.041	(0.029)
Risk					-0.011	(0.034)
Trust					0.212	(0.131)
Constant	1.056***	(0.204)	0.205	(0.401)	-0.585	(0.654)
Adj. R-squared			0.271		0.313	
N	175		144		114	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.2.1c: Regressions part 1 based on data set (iv)

	Ordered probit - dependent variable: number of games won in part 1					
	(1)		(2)		(3)	
Grade 1	-1.345***	(0.465)				
Grade 4	-0.069	(0.407)	-0.166	(0.409)		
Grade 9	-0.105	(0.401)	-0.059	(0.403)	-0.067	(0.408)
Uni young	0.251	(0.465)	0.143	(0.47)	0.143	(0.499)
Uni old	0.651	(0.465)	0.607	(0.468)	0.574	(0.499)
Grade 1 x female	0.116	(0.436)				
Grade 4 x female	-0.606	(0.413)	-0.381	(0.416)		
Grade 6 x female	-0.289	(0.437)	-0.187	(0.448)	-0.208	(0.462)
Grade 9 x female	-0.671	(0.461)	-0.616	(0.460)	-0.679	(0.467)
Uni young x female	-0.050	(0.482)	-0.082	(0.486)	-0.106	(0.511)
Uni old x female	-0.785	(0.499)	-0.832	(0.507)	-0.840	(0.536)
Uncover treasure	0.693***	(0.206)	0.712***	(0.227)	0.664**	(0.263)
Mark German			-0.026	(0.054)	-0.023	(0.062)
Mark math			0.137***	(0.046)	0.140***	(0.052)
Fairness					-0.179	(0.285)
Patience					-0.018	(0.051)
Risk					0.000	(0.058)
Trust					0.194	(0.221)
Pseudo R-squared	0.122		0.105		0.106	
N	146		116		88	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.2.2: Regressions part 1 (ordered probit) based on data set (i)

Online Appendix E.3: Number of games won in part 2

The OLS-regressions of analysis (i) in the paper that take the number of games won in part 2 as a dependent variable are based on data set (i). Below we present the OLS-regressions for the remaining three data sets in Tables E.3.1a, E.3.1b, and E.3.1c. As an additional robustness check we also present an ordered probit regression based on data set (i) in Table E.3.2.

	OLS - dependent variable: number of games won in part 2					
	(1)		(2)		(3)	
Grade 1	-0.935**	(0.394)				
Grade 4	-0.071	(0.357)	-0.129	(0.368)		
Grade 9	-0.180	(0.352)	-0.140	(0.363)	-0.143	(0.393)
Uni young	0.047	(0.406)	-0.040	(0.420)	0.017	(0.479)
Uni old	0.531	(0.404)	0.498	(0.418)	0.511	(0.476)
Grade 1 x female	0.033	(0.372)				
Grade 4 x female	-1.051***	(0.358)	-0.911**	(0.375)		
Grade 6 x female	-0.848**	(0.382)	-0.845**	(0.404)	-0.872*	(0.445)
Grade 9 x female	0.203	(0.399)	0.240	(0.411)	0.112	(0.448)
Uni young x female	-0.168	(0.420)	-0.156	(0.435)	-0.126	(0.488)
Uni old x female	-0.130	(0.434)	-0.160	(0.452)	-0.165	(0.511)
Uncover treasure	0.358**	(0.174)	0.450**	(0.199)	0.501**	(0.247)
Mark German			-0.020	(0.049)	-0.014	(0.060)
Mark math			0.097**	(0.040)	0.102**	(0.049)
Fairness					-0.280	(0.274)
Patience					0.004	(0.049)
Risk					0.017	(0.056)
Trust					0.420*	(0.212)
Constant	1.413***	(0.276)	0.634	(0.575)	-0.610	(1.078)
Adj. R-squared	0.267		0.195		0.259	
N	146		116		88	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.3.1a: Regressions part 2 based on data set (ii)

	OLS - dependent variable: number of games won in part 2					
	(1)		(2)		(3)	
Grade 1	-0.912**	(0.375)				
Grade 4	0.011	(0.333)	-0.011	(0.339)		
Grade 9	-0.242	(0.324)	-0.182	(0.331)	-0.166	(0.352)
Uni young	0.202	(0.348)	0.148	(0.355)	0.235	(0.388)
Uni old	0.494	(0.339)	0.463	(0.346)	0.483	(0.378)
Grade 1 x female	-0.077	(0.362)				
Grade 4 x female	-1.149***	(0.340)	-1.049***	(0.350)		
Grade 6 x female	-0.876**	(0.341)	-0.778**	(0.358)	-0.783**	(0.387)
Grade 9 x female	0.280	(0.342)	0.301	(0.347)	0.278	(0.369)
Uni young x female	-0.065	(0.361)	-0.071	(0.371)	-0.084	(0.398)
Uni old x female	-0.058	(0.351)	-0.085	(0.362)	-0.058	(0.404)
Uncover treasure	0.409	(0.153)	0.441**	(0.170)	0.520**	(0.203)
Mark German			0.008	(0.041)	0.015	(0.048)
Mark math			0.068**	(0.033)	0.068*	(0.038)
Fairness					-0.189	(0.225)
Patience					0.011	(0.038)
Risk					0.004	(0.045)
Trust					0.333*	(0.173)
Constant	2.344***	(0.254)	1.569***	(0.493)	0.536	(0.861)
Adj. R-squared	0.304		0.294		0.262	
N	175		144		114	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.3.1b: Regressions part 2 based on data set (iii)

	OLS - dependent variable: number of games won in part 2					
	(1)		(2)		(3)	
Grade 1	-0.911**	(0.364)				
Grade 4	0.010	(0.324)	-0.010	(0.330)		
Grade 9	-0.125	(0.315)	-0.066	(0.323)	-0.052	(0.341)
Uni young	0.203	(0.338)	0.150	(0.346)	0.241	(0.375)
Uni old	0.493	(0.330)	0.467	(0.338)	0.482	(0.366)
Grade 1 x female	0.022	(0.351)				
Grade 4 x female	-1.148***	(0.330)	-1.041***	(0.342)		
Grade 6 x female	-0.877***	(0.331)	-0.774**	(0.349)	-0.785**	(0.375)
Grade 9 x female	0.162	(0.332)	0.183	(0.338)	0.160	(0.357)
Uni young x female	-0.066	(0.350)	-0.062	(0.361)	-0.084	(0.386)
Uni old x female	-0.058	(0.341)	-0.076	(0.353)	-0.050	(0.391)
Uncover treasure	0.405***	(0.149)	0.444***	(0.166)	0.528***	(0.197)
Mark German			0.001	(0.040)	0.008	(0.046)
Mark math			0.072**	(0.032)	0.073*	(0.037)
Fairness					-0.218	(0.218)
Patience					0.017	(0.037)
Risk					0.007	(0.043)
Trust					0.340**	(0.167)
Constant	1.346***	(0.247)	0.594	(0.481)	-0.491	(0.835)
Adj. R-squared	0.306		0.303		0.273	
N	175		144		114	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.3.1c: Regressions part 2 based on data set (iv)

	Ordered probit - dependent variable: number of games won in part 2					
	(1)		(2)		(3)	
Grade 1	-1.070**	(0.450)	0.000			
Grade 4	-0.086	(0.399)	-0.144	(0.400)		
Grade 9	-0.331	(0.395)	-0.304	(0.396)	-0.283	(0.398)
Uni young	0.052	(0.453)	-0.031	(0.457)	0.008	(0.484)
Uni old	0.510	(0.458)	0.493	(0.461)	0.488	(0.486)
Grade 1 x female	-0.123	(0.424)				
Grade 4 x female	-1.223***	(0.415)	-1.074**	(0.423)		
Grade 6 x female	-0.985**	(0.437)	-1.003**	(0.453)	-0.953**	(0.461)
Grade 9 x female	0.349	(0.445)	0.397	(0.447)	0.202	(0.452)
Uni young x female	-0.197	(0.470)	-0.190	(0.473)	-0.113	(0.493)
Uni old x female	-0.111	(0.487)	-0.148	(0.494)	-0.094	(0.518)
Uncover treasure	0.386*	(0.198)	0.448**	(0.221)	0.460*	(0.253)
Mark German			-0.020	(0.054)	-0.015	(0.061)
Mark math			0.097**	(0.044)	0.095*	(0.050)
Fairness					-0.286	(0.278)
Patience					-0.006	(0.049)
Risk					0.029	(0.057)
Trust					0.466**	(0.219)
Pseudo R-squared	0.103		0.100		0.089	
N	146		116		88	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.3.2: Regressions part 2 (ordered probit) based on data set (i)

Online Appendix E.4: Number of games won in total

In the regressions we control for several characteristics of the subjects. In the OLS-regressions presented in Table 5 (specification (3)) of the paper and in Appendix E.3 the performance within part 2 is positively correlated with math grades and self-reported trust. In the main part of the paper we point out that this positive relationship is also found when considering the number of games won in total as a dependent variable (i.e. in part 1 *and* part 2). The respective regression results are shown in Table E.4.1.

The weakly significant gender effect observed for sixth graders reported in the paper for part 2 turns insignificant for data set (i) when pooling observations from part 1 and part 2. However, it remains significant or weakly significant when running the regression on data sets (ii), (iii), and (iv) (not shown). As an additional robustness check we also present an ordered probit regression based on data set (i) in Table E.4.2.

	OLS - dependent variable: number of games won in total					
	(1)		(2)		(3)	
Grade 1	-1.899***	(0.617)				
Grade 4	-0.143	(0.559)	-0.255	(0.581)		
Grade 9	-0.306	(0.552)	-0.239	(0.573)	-0.264	(0.601)
Uni young	0.233	(0.636)	0.080	(0.663)	0.094	(0.733)
Uni old	0.991	(0.633)	0.936	(0.660)	0.894	(0.728)
Grade 1 x female	0.017	(0.583)				
Grade 4 x female	-1.475**	(0.561)	-1.194*	(0.591)		
Grade 6 x female	-1.084*	(0.598)	-0.982	(0.638)	-0.974	(0.681)
Grade 9 x female	-0.223	(0.625)	-0.160	(0.649)	-0.347	(0.685)
Uni young x female	-0.219	(0.658)	-0.209	(0.686)	-0.148	(0.747)
Uni old x female	-0.707	(0.680)	-0.760	(0.714)	-0.717	(0.781)
Uncover treasure	0.828***	(0.272)	0.955***	(0.315)	0.973**	(0.378)
Mark German			-0.041	(0.077)	-0.034	(0.092)
Mark math			0.188***	(0.063)	0.195**	(0.075)
Fairness					-0.394	(0.419)
Patience					-0.023	(0.075)
Risk					0.027	(0.086)
Trust					0.570*	(0.324)
Constant	4.253***	(0.432)	2.798***	(0.908)	1.303	(1.648)
Adj. R-squared	0.260		0.211		0.151	
N	146		116		88	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.4.1: Regressions pooled based on data set (i)

	Ordered probit - dependent variable: number of games won					
	(1)		(2)		(3)	
Grade 1	-1.312***	(0.436)				
Grade 4	-0.094	(0.387)	-0.171	(0.388)		
Grade 9	-0.286	(0.383)	-0.271	(0.385)	-0.271	(0.388)
Uni young	0.148	(0.442)	0.062	(0.446)	0.080	(0.475)
Uni old	0.709	(0.447)	0.720	(0.450)	0.708	(0.478)
Grade 1 x female	-0.056	(0.408)				
Grade 4 x female	-1.068***	(0.397)	-0.849**	(0.404)		
Grade 6 x female	-0.703*	(0.418)	-0.642	(0.429)	-0.643	(0.441)
Grade 9 x female	-0.072	(0.435)	-0.002	(0.438)	-0.150	(0.443)
Uni young x female	-0.137	(0.458)	-0.159	(0.462)	-0.122	(0.484)
Uni old x female	-0.542	(0.477)	-0.599	(0.484)	-0.558	(0.510)
Uncover treasure	0.611***	(0.194)	0.668***	(0.217)	0.684***	(0.252)
Mark German			-0.034	(0.052)	-0.032	(0.059)
Mark math			0.136***	(0.044)	0.139***	(0.050)
Fairness					-0.277	(0.272)
Patience					-0.009	(0.048)
Risk					0.025	(0.055)
Trust					0.395*	(0.212)
Pseudo R-squared	0.099		0.091		0.089	
N	146		116		88	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.4.2: Regressions pooled (ordered probit) based on data set (i)

Online Appendix E.5: Difference in the number of games won between part 1 and part 2

Also the regressions of analysis (ii) on the improvement between part 1 and part 2 with respect to Hypothesis 2 in the paper are based on data set (i). Below we present these OLS-regressions for the remaining three data sets with the number of games won in part 2 minus the number of games won in part 1 as the dependent variable in Tables E.5.1a, E.5.1b, and E.5.1c. As an additional robustness check we also present an ordered probit regression based on data set (i) in Table E.5.2.

	OLS - dependent variable: difference in number of games won between part 1 and part 2					
	(1)		(2)		(3)	
Grade 1	0.070	(0.376)				
Grade 4	0.071	(0.341)	0.070	(0.348)		
Grade 9	-0.161	(0.336)	-0.154	(0.344)	-0.146	(0.38)
Uni young	-0.096	(0.387)	-0.103	(0.398)	-0.071	(0.463)
Uni old	0.126	(0.386)	0.127	(0.395)	0.141	(0.460)
Grade 1 x female	-0.130	(0.355)				
Grade 4 x female	-0.715**	(0.342)	-0.709**	(0.354)		
Grade 6 x female	-0.609*	(0.365)	-0.617	(0.382)	-0.630	(0.430)
Grade 9 x female	0.623	(0.381)	0.628	(0.389)	0.533	(0.433)
Uni young x female	-0.082	(0.401)	-0.060	(0.411)	-0.016	(0.472)
Uni old x female	0.444	(0.414)	0.457	(0.428)	0.455	(0.493)
Uncover treasure	0.005	(0.166)	0.052	(0.189)	0.116	(0.239)
Mark German			-0.011	(0.046)	0.000	(0.058)
Mark math			0.006	(0.038)	0.012	(0.047)
Fairness					-0.176	(0.265)
Patience					0.005	(0.047)
Risk					0.018	(0.054)
Trust					0.324	(0.205)
Constant	0.425	(0.264)	0.451	(0.544)	-0.677	(1.041)
Adj. R-squared	0.047		0.052		-0.010	
N	146		116		88	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.5.1a: Regressions on improvement based on data set (ii)

	OLS - dependent variable: difference in number of games won between part 1 and part 2					
	(1)		(2)		(3)	
Grade 1	0.150	(0.396)				
Grade 4	0.148	(0.352)	0.157	(0.359)		
Grade 9	-0.106	(0.342)	-0.119	(0.351)	-0.125	(0.385)
Uni young	-0.040	(0.367)	-0.031	(0.376)	-0.109	(0.424)
Uni old	-0.031	(0.358)	-0.022	(0.368)	-0.094	(0.413)
Grade 1 x female	-0.185	(0.382)				
Grade 4 x female	-0.644*	(0.359)	-0.687*	(0.372)		
Grade 6 x female	-0.334	(0.360)	-0.443	(0.380)	-0.389	(0.423)
Grade 9 x female	0.927**	(0.361)	0.922**	(0.368)	0.891**	(0.403)
Uni young x female	0.080	(0.381)	0.084	(0.393)	0.101	(0.436)
Uni old x female	0.588	(0.370)	0.585	(0.384)	0.636	(0.442)
Uncover treasure	-0.116	(0.162)	-0.102	(0.180)	-0.047	(0.222)
Mark German			0.003	(0.043)	0.003	(0.052)
Mark math			-0.025	(0.035)	-0.028	(0.042)
Fairness					-0.192	(0.247)
Patience					-0.024	(0.042)
Risk					0.020	(0.049)
Trust					0.059	(0.189)
Constant	0.455*	(0.269)	0.663	(0.523)	0.673	(0.942)
Adj. R-squared	0.048		0.057		0.002	
N	175		144		114	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.5.1b: Regressions on improvement based on data set (iii)

	OLS - dependent variable: difference in number of games won between part 1 and part 2					
	(1)		(2)		(3)	
Grade 1	0.181	(0.368)				
Grade 4	0.224	(0.327)	0.230	(0.334)		
Grade 9	-0.070	(0.318)	-0.072	(0.326)	-0.066	(0.355)
Uni young	-0.012	(0.341)	-0.012	(0.349)	-0.080	(0.391)
Uni old	0.048	(0.333)	0.057	(0.341)	-0.015	(0.381)
Grade 1 x female	-0.140	(0.355)				
Grade 4 x female	-0.740**	(0.333)	-0.764**	(0.345)		
Grade 6 x female	-0.447	(0.335)	-0.474	(0.353)	-0.405	(0.390)
Grade 9 x female	0.767**	(0.335)	0.767**	(0.342)	0.734*	(0.372)
Uni young x female	0.132	(0.354)	0.145	(0.365)	0.151	(0.402)
Uni old x female	0.567	(0.344)	0.565	(0.356)	0.604	(0.408)
Uncover treasure	0.032	(0.150)	0.063	(0.167)	0.128	(0.205)
Mark German			-0.001	(0.040)	0.007	(0.048)
Mark math			-0.012	(0.033)	-0.010	(0.039)
Fairness					-0.224	(0.228)
Patience					-0.024	(0.038)
Risk					0.018	(0.045)
Trust					0.128	(0.174)
Cut 1	0.290	(0.25)	0.390	(0.486)	0.094	(0.869)
Adj. R-squared	0.061		0.067		0.014	
N	175		144		114	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.5.1c: Regressions on improvement based on data set (iv)

	Ordered Probit - dependent variable: difference in number of games won between part 1 and part 2					
	(1)		(2)		(3)	
Grade 1	0.050	(0.440)				
Grade 4	0.035	(0.401)	0.052	(0.402)		
Grade 9	-0.148	(0.395)	-0.129	(0.396)	-0.107	(0.398)
Uni young	-0.125	(0.455)	-0.119	(0.459)	-0.019	(0.485)
Uni old	0.065	(0.453)	0.053	(0.455)	0.112	(0.481)
Grade 1 x female	-0.186	(0.416)				
Grade 4 x female	-0.761*	(0.407)	-0.784*	(0.415)		
Grade 6 x female	-0.721*	(0.434)	-0.841*	(0.448)	-0.843*	(0.458)
Grade 9 x female	0.797*	(0.446)	0.784*	(0.447)	0.649	(0.452)
Uni young x female	-0.159	(0.471)	-0.154	(0.475)	-0.130	(0.493)
Uni old x female	0.500	(0.487)	0.499	(0.494)	0.447	(0.516)
Uncover treasure	-0.136	(0.195)	-0.096	(0.218)	-0.014	(0.250)
Mark German			-0.001	(0.053)	0.003	(0.060)
Mark math			-0.003	(0.044)	-0.001	(0.049)
Fairness					-0.166	(0.277)
Patience					0.025	(0.050)
Risk					0.011	(0.056)
Trust					0.303	(0.215)
Adj. R-squared	0.045		0.056		0.053	
N	146		116		88	

Standard errors are given in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.5.2: Regressions on improvement (ordered probit) based on data set (i)

Online Appendix E.6: Correctly identified winning positions in part 1

In order to maximize the likelihood function of analysis (iii) presented in section 5.4 some distribution for the random variable *T* has to be selected. The positive random variable *T* can be thought of as the number of correctly identified winning positions in the remaining game or as the number of successful steps of backward reasoning. Table E.6.1 shows the Bayesian information criterion (BIC) values for part 1 that result from different distributional assumptions typically made in survival analysis. The BIC is an indicator for model fit punishing the use of additional parameters. Based on BIC we pick the generalized gamma distribution as it yields the lowest value across model specifications.

Tables E.6.2.a and E.6.2b show the results for the regressions based on data sets (iii) and (iv) based on interval-censoring and the generalized gamma distribution. The models for data set (ii) did not achieve convergence and could not be estimated using the same procedure. Also specification (3) based on data set (iv) did not achieve convergence.

In the analysis of correctly identified winning positions we do not count moves as "correct" that were made to a winning position, but cannot be based on backward analysis. These moves were made before uncovering the treasure and therefore made without knowing its position. Table E.6.3 shows the regressions based on data set (i) that result when counting these moves as correct anyway.

We also considered two alternative approaches that yielded similar results: First, the method introduced by Royston and Parmar (2002) that models the baseline cumulative hazard function of a proportional hazards model as a cubic spline (see Table E.6.4). Second, a Cox proportional hazards model with right-censoring and midpoint imputation (see Table E.6.5).

	Part 1		
Distribution	(1)	(2)	(3)
Exponential	1578.801	1272.773	994.707
Weibull	1578.793	1263.760	985.327
Gompertz	1585.956	1277.613	998.080
Log-logistic	1540.285	1234.048	963.439
Log-normal	1538.458	1231.163	960.794
Gamma	1571.521	1255.151	978.208
Generalized gamma	1489.294	1199.871	937.271
Inverse Gaussian	1535.972	1227.410	957.909

Table E.6.1: Model comparisons based on data set (i): Bayesian information criterion

	Survival analysis - dependent variable: correctly identified winning positions					
	(1)		(2)		(3)	
Grade 1	0.806	(0.206)				
Grade 4	0.957	(0.764)	0.898	(0.493)		
Grade 9	0.841	(0.289)	0.819	(0.253)	0.824	(0.231)
Uni young	1.359	(0.108)	1.237	(0.228)	1.352*	(0.091)
Uni old	1.385*	(0.066)	1.269	(0.148)	1.397*	(0.076)
Grade 1 x female	0.830	(0.125)				
Grade 4 x female	0.845	(0.280)	0.874	(0.401)		
Grade 6 x female	0.729*	(0.080)	0.785	(0.156)	0.786	(0.197)
Grade 9 x female	0.794	(0.195)	0.790	(0.170)	0.834	(0.338)
Uni young x female	0.801	(0.118)	0.817	(0.155)	0.814	(0.132)
Uni old x female	0.734**	(0.012)	0.761**	(0.035)	0.723**	(0.018)
Mark German			0.996	(0.814)	1.001	(0.973)
Mark math			1.040**	(0.018)	1.040**	(0.021)
Fairness					1.022	(0.801)
Patience					1.009	(0.538)
Risk					0.981	(0.282)
Trust					1.130*	(0.086)
Constant	1.548**	(0.028)	1.197	(0.495)	0.757	(0.482)
Log pseudolikelihood	-827.598		-679.754		-535.684	
Number of moves	1586		1334		1068	
Number of individuals	175		144		114	

Robust standard errors clustered on the individual level are given in parentheses. Coefficients of the accelerated failure time model are exponentiated. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.6.2a: Regressions part 1 based on data set (iii)

	Survival analysis - dependent variable: correctly identified winning positions					
	(1)		(2)		(3)	
Grade 1	0.615***	(0.008)			-	-
Grade 4	0.717	(0.319)	0.714	(0.117)	-	-
Grade 9	0.826	(0.327)	0.858	(0.355)	-	-
Uni young	1.163	(0.322)	1.104	(0.509)	-	-
Uni old	1.195	(0.201)	1.140	(0.331)	-	-
Grade 1 x female	0.844	(0.455)			-	-
Grade 4 x female	0.971	(0.939)	0.990	(0.964)	-	-
Grade 6 x female	0.707**	(0.010)	0.728**	(0.030)	-	-
Grade 9 x female	0.766	(0.262)	0.732*	(0.057)	-	-
Uni young x female	0.797	(0.117)	0.804	(0.137)	-	-
Uni old x female	0.697**	(0.027)	0.732**	(0.040)	-	-
Mark German			1.000	(0.992)	-	-
Mark math			1.036**	(0.035)	-	-
Fairness					-	-
Patience					-	-
Risk					-	-
Trust					-	-
Constant	1.640	(0.162)	1.259	(0.329)	-	-
Log pseudolikelihood	-782.052		-646.425		-	
Number of moves	1411		1190		-	
Number of individuals	175		144		-	

Robust standard errors clustered on the individual level are given in parentheses. Coefficients of the accelerated failure time model are exponentiated. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.6.2b: Regressions part 1 based on data set (iv)

	Survival analysis - dependent variable: correctly identified winning positions					
	(1)		(2)		(3)	
Grade 1	0.857	(0.166)				
Grade 4	0.991	(0.105)	0.944	(0.057)		
Grade 9	0.921	(0.123)	0.906	(0.067)	0.913	(0.121)
Uni young	1.319	(0.282)	1.282	(0.227)	1.269	(0.253)
Uni old	1.475**	(0.280)	1.446**	(0.227)	1.455**	(0.267)
Grade 1 x female	0.883	(0.135)				
Grade 4 x female	0.917	(0.132)	1.015	(0.121)		
Grade 6 x female	0.928	(0.130)	1.059	(0.097)	1.073	(0.117)
Grade 9 x female	0.839	(0.139)	0.879	(0.144)	0.867	(0.103)
Uni young x female	0.872	(0.176)	0.882	(0.193)	0.914	(0.192)
Uni old x female	0.715**	(0.119)	0.739*	(0.119)	0.707**	(0.117)
Mark German			0.977	(0.014)	0.985	(0.024)
Mark math			1.025	(0.02)	1.027	(0.020)
Fairness					0.975	(0.067)
Patience					0.992	(0.009)
Risk					0.994	(0.017)
Trust					1.085	(0.089)
Constant	1.220	(0.302)	1.005	(0.123)	0.798	(0.354)
Log pseudolikelihood	-723.690		-570.794		-430.039	
Number of moves	1,317		1,073		821	
Number of individuals	146		116		88	

Robust standard errors clustered on the individual level are given in parentheses. Coefficients of the accelerated failure time model are exponentiated. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.6.3: Regressions part 1 based on data set (i) - without controlling for uncovering

	Survival analysis - dependent variable: correctly identified winning positions					
	(1)		(2)		(3)	
Grade 1	0.360*	(0.186)				
Grade 4	0.026	(0.267)	0.037	(0.229)		
Grade 9	0.298	(0.226)	0.302	(0.220)	0.286	(0.243)
Uni young	-0.139	(0.256)	-0.125	(0.312)	-0.251	(0.332)
Uni old	-0.213	(0.245)	-0.207	(0.242)	-0.284	(0.252)
Grade 1 x female	0.188	(0.172)				
Grade 4 x female	0.419	(0.283)	0.391*	(0.226)		
Grade 6 x female	0.408*	(0.231)	0.343	(0.245)	0.395	(0.290)
Grade 9 x female	0.282	(0.280)	0.284	(0.268)	0.271	(0.321)
Uni young x female	0.173	(0.314)	0.165	(0.304)	0.228	(0.279)
Uni old x female	0.467**	(0.216)	0.458**	(0.229)	0.528**	(0.220)
Mark German			0.008	(0.027)	0.008	(0.034)
Mark math			-0.016	(0.026)	0.003	(0.032)
Fairness					0.059	(0.152)
Patience					-0.030	(0.029)
Risk					0.009	(0.037)
Trust					-0.009	(0.127)
Constant	-2.959***	(0.285)	-3.266***	(0.447)	-3.442***	(0.869)
Log pseudolikelihood	-677.188		-538.531		-407.780	
Number of moves	1,317		1,073		821	
Number of individuals	146		116		88	

Bootstrapped standard errors clustered on the individual level (100 draws) are given in parentheses. Parameters are given as log-hazard ratios. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.6.4: Royston-Parmar regressions part 1 based on data set (i)

	Survival analysis - dependent variable: correctly identified winning positions					
	(1)		(2)		(3)	
Grade 1	1.389**	(0.226)				
Grade 4	1.010	(0.193)	1.032	(0.195)		
Grade 9	1.237	(0.232)	1.234	(0.237)	1.223	(0.233)
Uni young	0.871	(0.202)	0.888	(0.208)	0.824	(0.207)
Uni old	0.817	(0.159)	0.822	(0.167)	0.781	(0.168)
Grade 1 x female	1.136	(0.134)				
Grade 4 x female	1.392*	(0.258)	1.345	(0.244)		
Grade 6 x female	1.354*	(0.235)	1.293	(0.237)	1.330	(0.237)
Grade 9 x female	1.271	(0.226)	1.276	(0.218)	1.278	(0.226)
Uni young x female	1.166	(0.263)	1.164	(0.266)	1.189	(0.280)
Uni old x female	1.440**	(0.248)	1.446**	(0.255)	1.487**	(0.258)
Mark German			1.005	(0.021)	1.005	(0.023)
Mark math			0.978	(0.017)	0.991	(0.020)
Fairness					1.035	(0.124)
Patience					0.984	(0.019)
Risk					1.002	(0.024)
Trust					0.968	(0.083)
Log pseudolikelihood	-4188.778		-3097.839		-2224.270	
Number of moves	1,317		1,073		821	
Number of individuals	146		116		88	

Robust standard errors clustered on the individual level are given in parentheses. Parameters are given as hazard ratios. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.6.5: Cox proportional hazard regressions with midpoint imputation part 1 based on data set (i)

Online Appendix E.7: Correctly identified winning positions in part 2

Table E.7.1 shows the Bayesian information criterion (BIC) values for part 2 that result from different distributional assumptions typically made in survival analysis. Also for part 2 the generalized gamma distribution yields the best model fit.

Tables E.7.2.a, E.7.2b, and E.7.2c show the results for the regressions based on data sets (ii), (iii), and (iv) based on interval-censoring and the generalized gamma distribution. Table E.7.3 shows the regressions that result when counting moves as correct that were made to winning positions before uncovering the treasure for data set (i).

Table E.7.4 shows the regressions based on data set (i) using the method introduced by Royston and Parmar (2002). A Cox proportional hazards model with right-censoring and midpoint imputation for data set (i) is presented in Table E.7.5.

	Part 2		
Distribution	(1)	(2)	(3)
Exponential	1598.070	1235.738	959.325
Weibull	1572.971	1194.946	915.244
Gompertz	1596.229	1221.798	937.968
Log-logistic	1530.375	1160.217	892.869
Log-normal	1525.277	1156.591	890.278
Gamma	1557.557	1178.951	903.286
Generalized gamma	1490.304	1141.926	886.949
Inverse Gaussian	1527.450	1157.128	889.196

Table E.7.1: Model comparisons based on data set (i): Bayesian information criterion

	Survival analysis - dependent variable: correctly identified winning positions					
	(1)		(2)		(3)	
Grade 1	0.683*	(0.087)				
Grade 4	1.000	(0.999)	0.993	(0.961)		
Grade 9	0.910	(0.566)	0.929	(0.656)	0.919	(0.571)
Uni young	1.094	(0.640)	1.067	(0.719)	1.103	(0.564)
Uni old	1.304	(0.207)	1.284	(0.243)	1.291	(0.216)
Grade 1 x female	0.898	(0.662)				
Grade 4 x female	0.574***	(0.002)	0.624***	(0.003)		
Grade 6 x female	0.581***	(0.004)	0.598***	(0.005)	0.603***	(0.003)
Grade 9 x female	0.913	(0.634)	0.946	(0.749)	0.879	(0.446)
Uni young x female	0.892	(0.537)	0.885	(0.473)	0.911	(0.566)
Uni old x female	0.808	(0.442)	0.839	(0.477)	0.918	(0.696)
Mark German			0.998	(0.915)	0.997	(0.900)
Mark math			1.039*	(0.055)	1.036*	(0.077)
Fairness					0.911	(0.380)
Patience					1.016	(0.405)
Risk					1.020	(0.315)
Trust					1.196**	(0.031)
Constant	2.291***	(0.000)	1.757*	(0.077)	1.020	(0.968)
Log pseudolikelihood	-679.804		-512.948		-382.160	
Number of moves	1278		1020		792	
Number of individuals	146		116		88	

Robust standard errors clustered on the individual level are given in parentheses. Coefficients of the accelerated failure time model are exponentiated. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.7.2a: Regressions part 2 based on data set (ii)

	Survival analysis - dependent variable: correctly identified winning positions					
	(1)		(2)		(3)	
Grade 1	0.723*	(0.064)				
Grade 4	1.014	(0.911)	1.017	(0.890)		
Grade 9	0.776	(0.129)	0.814	(0.229)	0.816	(0.214)
Uni young	1.203	(0.196)	1.177	(0.241)	1.221	(0.137)
Uni old	1.255	(0.163)	1.233	(0.201)	1.239	(0.170)
Grade 1 x female	0.829	(0.322)				
Grade 4 x female	0.627***	(0.000)	0.647***	(0.000)		
Grade 6 x female	0.628***	(0.001)	0.657***	(0.003)	0.654***	(0.002)
Grade 9 x female	1.136	(0.446)	1.144	(0.404)	1.116	(0.500)
Uni young x female	0.910	(0.541)	0.900	(0.474)	0.899	(0.467)
Uni old x female	0.866	(0.493)	0.890	(0.538)	0.967	(0.846)
Mark German			1.010	(0.586)	1.007	(0.701)
Mark math			1.024	(0.113)	1.024	(0.116)
Fairness					0.971	(0.727)
Patience					1.007	(0.647)
Risk					1.010	(0.506)
Trust					1.118*	(0.076)
Constant	2.319***	(0.000)	1.800**	(0.017)	1.284	(0.458)
Log pseudolikelihood	-824.6853		-647.334		-510.088	
Number of moves	1720		1423		1150	
Number of individuals	175		144		114	

Robust standard errors clustered on the individual level are given in parentheses. Coefficients of the accelerated failure time model are exponentiated. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.7.2b: Regressions part 2 based on data set (iii)

	Survival analysis - dependent variable: correctly identified winning positions					
	(1)		(2)		(3)	
Grade 1	0.690*	(0.08)				
Grade 4	1.015	(0.910)	1.018	(0.888)		
Grade 9	0.881	(0.407)	0.918	(0.584)	0.905	(0.495)
Uni young	1.209	(0.201)	1.181	(0.247)	1.235	(0.120)
Uni old	1.262	(0.167)	1.236	(0.210)	1.233	(0.183)
Grade 1 x female	0.863	(0.543)				
Grade 4 x female	0.558***	(0.001)	0.597***	(0.001)		
Grade 6 x female	0.560***	(0.001)	0.604***	(0.004)	0.600***	(0.002)
Grade 9 x female	0.992	(0.963)	1.018	(0.904)	0.998	(0.988)
Uni young x female	0.909	(0.542)	0.900	(0.485)	0.891	(0.441)
Uni old x female	0.863	(0.489)	0.881	(0.508)	0.955	(0.790)
Mark German			1.007	(0.676)	1.004	(0.829)
Mark math			1.033**	(0.039)	1.033**	(0.041)
Fairness					0.952	(0.562)
Patience					1.016	(0.354)
Risk					1.011	(0.497)
Trust					1.133*	(0.055)
Constant	2.286***	(0.000)	1.631*	(0.069)	1.086	(0.815)
Log pseudolikelihood	-805.677		-634.771		-498.523	
Number of moves	1545		1279		1036	
Number of individuals	175		144		114	

Robust standard errors clustered on the individual level are given in parentheses. Coefficients of the accelerated failure time model are exponentiated. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.7.2c: Regressions part 2 based on data set (iv)

	Survival analysis - dependent variable: correctly identified winning positions					
	(1)		(2)		(3)	
Grade 1	0.807	(0.119)				
Grade 4	1.044	(0.147)	1.042	(0.133)		
Grade 9	0.831	(0.146)	0.849	(0.152)	0.855	(0.144)
Uni young	1.094	(0.218)	1.080	(0.189)	1.099	(0.190)
Uni old	1.339	(0.276)	1.319	(0.269)	1.322	(0.265)
Grade 1 x female	0.868	(0.109)				
Grade 4 x female	0.612***	(0.084)	0.644***	(0.077)		
Grade 6 x female	0.719*	(0.124)	0.743*	(0.120)	0.742*	(0.125)
Grade 9 x female	1.237	(0.220)	1.246	(0.223)	1.148	(0.192)
Uni young x female	0.890	(0.170)	0.893	(0.152)	0.928	(0.155)
Uni old x female	0.762	(0.228)	0.844	(0.206)	0.912	(0.203)
Mark German			0.990	(0.021)	0.994	(0.024)
Mark math			1.027	(0.020)	1.026	(0.021)
Fairness					0.918	(0.099)
Patience					1.004	(0.020)
Risk					1.016	(0.021)
Trust					1.203**	(0.094)
Constant	2.056***	(0.327)	2.123***	(0.565)	1.222	(0.530)
Log pseudolikelihood	-718.495		-530.684		-397.81	
Number of moves	1,424		1,136		880	
Number of individuals	146		116		88	

Robust standard errors clustered on the individual level are given in parentheses. Coefficients of the accelerated failure time model are exponentiated. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.7.3: Regressions part 2 based on data set (i) - without controlling for uncovering

	Survival analysis - dependent variable: correctly identified winning positions					
	(1)		(2)		(3)	
Grade 1	0.273	(0.273)				
Grade 4	-0.164	(0.212)	-0.153	(0.216)		
Grade 9	0.089	(0.267)	0.077	(0.259)	0.084	(0.253)
Uni young	-0.161	(0.289)	-0.131	(0.355)	-0.203	(0.332)
Uni old	-0.695	(0.451)	-0.688	(0.458)	-0.779*	(0.468)
Grade 1 x female	0.061	(0.292)				
Grade 4 x female	0.675***	(0.217)	0.621***	(0.204)		
Grade 6 x female	0.533**	(0.239)	0.481*	(0.249)	0.450*	(0.249)
Grade 9 x female	0.035	(0.301)	0.034	(0.302)	0.121	(0.310)
Uni young x female	0.143	(0.295)	0.151	(0.351)	0.080	(0.321)
Uni old x female	0.208	(0.513)	0.253	(0.499)	0.178	(0.502)
Mark German			-0.005	(0.040)	0.017	(0.036)
Mark math			-0.022	(0.027)	-0.035	(0.037)
Fairness					0.082	(0.188)
Patience					-0.024	(0.032)
Risk					-0.049	(0.040)
Trust					-0.238*	(0.139)
Constant	-3.623***	(0.374)	-4.043***	(0.581)	-3.234***	(0.819)
Log pseudolikelihood	-685.769		-512.244		-384.229	
Number of moves	1,424		1,136		880	
Number of individuals	146		116		88	

Bootstrapped standard errors clustered on the individual level (100 draws) are given in parentheses. Parameters are given as log-hazard ratios. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.7.4 Royston-Parmar regressions part 2 based on data set (i)

	Survival analysis - dependent variable: correctly identified winning positions					
	(1)		(2)		(3)	
Grade 1	1.450	(0.346)				
Grade 4	0.949	(0.180)	0.979	(0.190)		
Grade 9	1.118	(0.231)	1.112	(0.242)	1.113	(0.230)
Uni young	0.890	(0.221)	0.929	(0.232)	0.881	(0.225)
Uni old	0.586	(0.220)	0.592	(0.231)	0.554	(0.223)
Grade 1 x female	1.060	(0.242)				
Grade 4 x female	1.760***	(0.270)	1.694***	(0.252)		
Grade 6 x female	1.713***	(0.309)	1.713***	(0.338)	1.689***	(0.307)
Grade 9 x female	1.084	(0.258)	1.082	(0.252)	1.172	(0.275)
Uni young x female	1.172	(0.278)	1.189	(0.279)	1.117	(0.275)
Uni old x female	1.145	(0.492)	1.201	(0.510)	1.115	(0.457)
Mark German			0.995	(0.030)	1.012	(0.034)
Mark math			0.969	(0.021)	0.961	(0.025)
Fairness					1.111	(0.165)
Patience					0.986	(0.025)
Risk					0.962	(0.028)
Trust					0.801**	(0.082)
Log pseudolikelihood	-3851.606		-2727.388		-1919.758	
Number of moves	1,424		1,136		880	
Number of individuals	146		116		88	

Robust standard errors clustered on the individual level are given in parentheses. Parameters are given as hazard ratios. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.7.5: Cox proportional hazard regressions with midpoint imputation part 2 based on data set (i)

Online Appendix E.8: Correctly identified winning positions in total

We also conduct the regressions on the number of correctly identified winning positions for the total number of moves, i.e. for all moves from part 1 *and* part 2. Table E.8.1 summarizes the BIC values that result from different distributional assumptions. It reveals that, again, the generalized gamma distribution yields the best fit across the three specifications. The regressions based on interval-censoring and the generalized gamma distribution for data set (i) are shown in Table E.8.2.

	Pooled		
Distribution	(1)	(2)	(3)
Exponential	3118.191	2452.571	1890.555
Weibull	3091.543	2401.603	1838.100
Gompertz	3122.217	2442.509	1873.961
Log-logistic	3007.852	2335.921	1790.667
Log-normal	3000.285	2328.224	1784.082
Gamma	3068.451	2376.750	1817.758
Generalized gamma	2904.551	2273.014	1749.414
Inverse Gaussian	3005.460	2329.600	1785.929

Table E.8.1: Model comparisons based on data set (i): Bayesian information criterion

	Survival analysis - dependent variable: correctly identified winning positions					
	(1)		(2)		(3)	
Grade 1	0.750**	(0.102)				
Grade 4	0.990	(0.131)	0.971	(0.121)		
Grade 9	0.838	(0.113)	0.830	(0.116)	0.833	(0.111)
Uni young	1.164	(0.198)	1.122	(0.179)	1.127	(0.182)
Uni old	1.307	(0.217)	1.266	(0.212)	1.254	(0.206)
Grade 1 x female	0.845	(0.116)				
Grade 4 x female	0.734**	(0.098)	0.752**	(0.099)		
Grade 6 x female	0.736**	(0.096)	0.753**	(0.096)	0.748**	(0.095)
Grade 9 x female	0.906	(0.142)	0.921	(0.143)	0.868	(0.132)
Uni young x female	0.889	(0.140)	0.889	(0.135)	0.920	(0.135)
Uni old x female	0.754	(0.156)	0.799	(0.154)	0.847	(0.151)
Mark German			0.991	(0.017)	0.993	(0.019)
Mark math			1.027*	(0.016)	1.023	(0.016)
Fairness					0.938	(0.075)
Patience					1.003	(0.015)
Risk					1.015	(0.017)
Trust					1.133*	(0.074)
Part 2	1.199**	(0.048)	1.196***	(0.053)	1.212***	(0.062)
Constant	1.745***	(0.207)	1.642**	(0.361)	1.135	(0.400)
Log pseudolikelihood	-1392.904		-1078.755		-811.476	
Number of moves	2,741		2,209		1,701	
Number of individuals	146		116		88	

Robust standard errors clustered on the individual level are given in parentheses. Coefficients of the accelerated failure time model are exponentiated. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table E.8.2: Regressions pooled based on data set (i)

FOR ONLINE PUBLICATION

Online Appendix F: Additional figures

This Appendix graphically presents the share of subjects winning by game length separately for the six age groups (see also Figure 4).

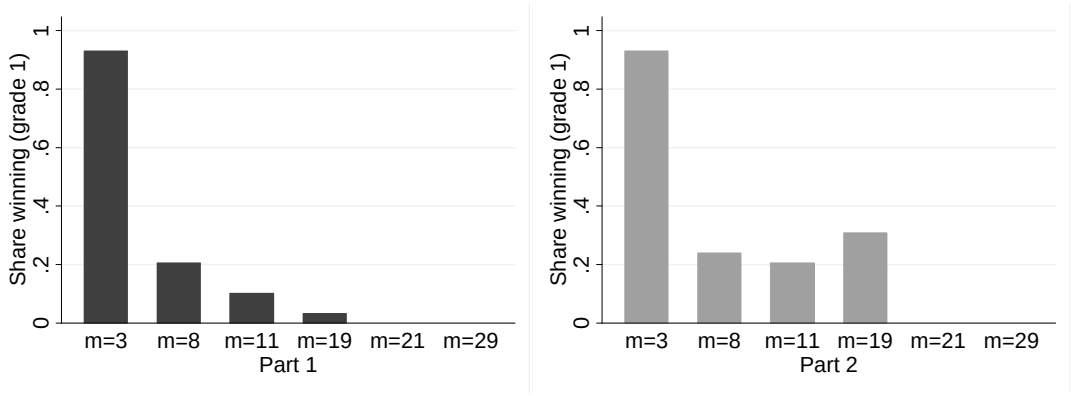


Figure F.1: Share of subjects winning by game length – grade 1

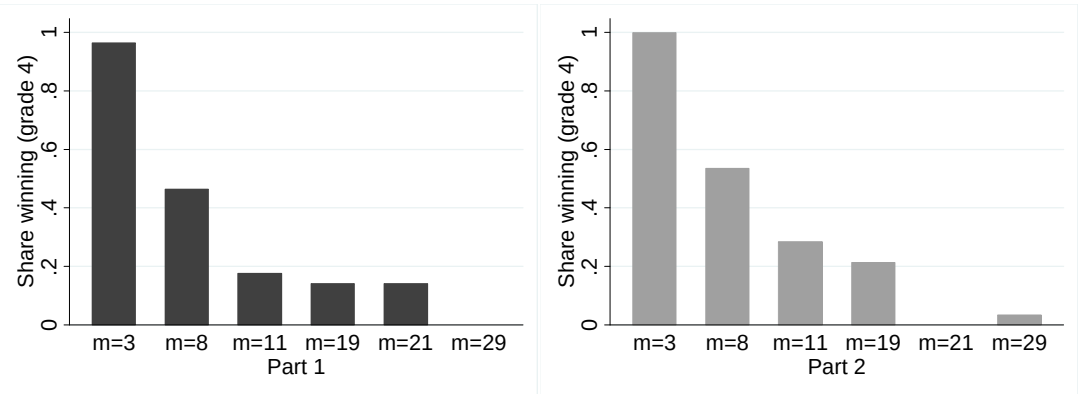


Figure F.2: Share of subjects winning by game length – grade 4

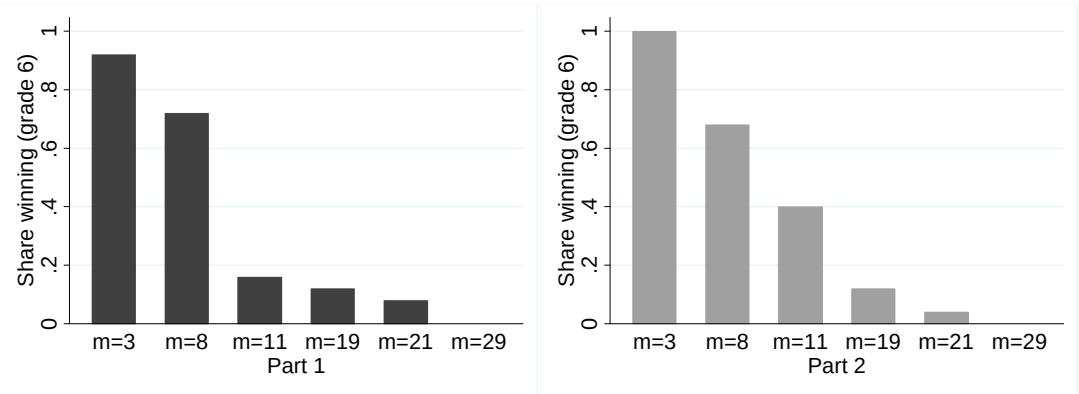


Figure F.3: Share of subjects winning by game length – grade 6

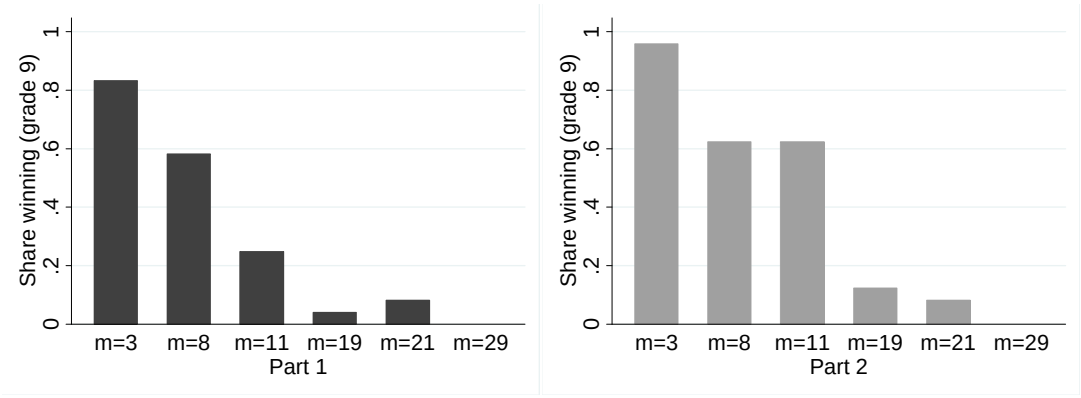


Figure F.4: Share of subjects winning by game length – grade 9

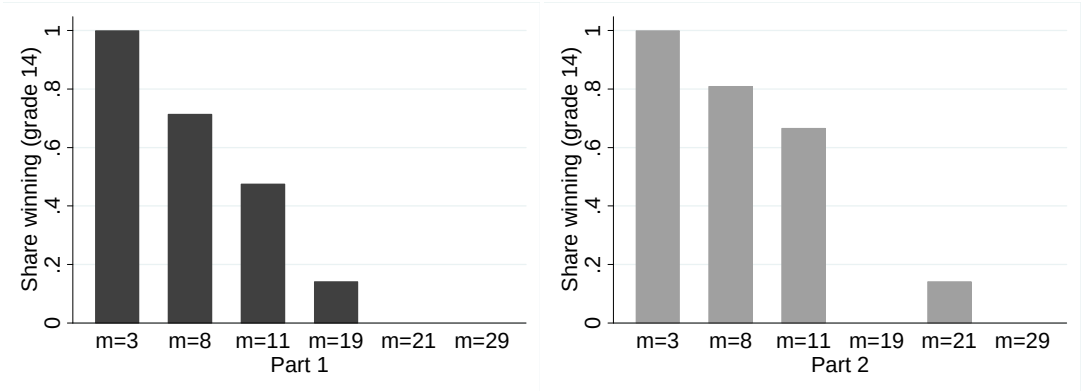


Figure F.5: Share of subjects winning by game length – uni young

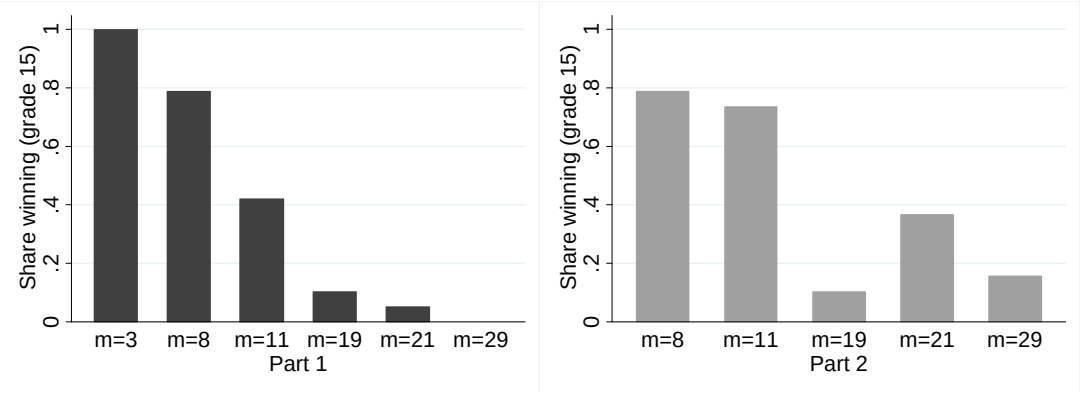


Figure F.6: Share of subjects winning by game length – uni old