Steps of reasoning in children and adolescents *

Isabelle Brocas
University of Southern California and CEPR

Juan D. Carrillo
University of Southern California and CEPR

November 2020

Abstract

We develop a novel graphical paradigm of a strict dominance solvable game to study the developmental trajectory of steps of reasoning between 8 years old and adulthood. Most participants play the equilibrium action either always or only when they have a dominant strategy. Although age is a determinant of equilibrium choice, some very young participants display an innate ability to play at equilibrium. Finally, the proportion of equilibrium play increases significantly until 5th grade and stabilizes afterwards, suggesting that the contribution of age to equilibrium play vanishes early in life.

Keywords: developmental decision-making, backward induction, steps of dominance.

JEL Classification: C91, D83.

*We thank Sobhana Atluri, the members of the Los Angeles Behavioral Economics Laboratory (LABEL) and seminar participants at USC and Stanford for their insights and comments, and Chris Crabbe for exceptional programming. We are grateful to the staff of the Lycée International de Los Angeles (LILA) and Thomas Starr King (KING) –in particular Emmanuelle Acker, Nordine Bouriche, Adriana Díaz, Mathieu Mondange and Anneli Harvey– for their help and support running the experiment in their schools. The study was conducted with the University of Southern California IRB approval UP-12-00528. We gratefully acknowledge the financial support of the National Science Foundation grant SES-1851915. Address for correspondence: Isabelle Brocas, Department of Economics, University of Southern California, 3620 S. Vermont Ave., Los Angeles, CA 90089, USA, <brocas@usc.edu>.

1 Introduction

At which age are individuals capable of selecting rational, forward-looking, optimal decisions in multi-person games of strategy? The goal of this paper is to provide an answer in a simple, well-defined game theoretic setting. Existing research on adults documents significant disparity in depths of reasoning across individuals,[1] and a positive correlation between cognitive skills and strategic sophistication in economic choices.[2] In this paper, we argue that a key to understand heterogeneity in strategic reasoning by adults is to unveil the process through which sophistication develops. In other words, observing how the ability to think strategically evolves with age should provide invaluable evidence to identify general patterns in the acquisition of sophistication, differences across individual trajectories, and causal mechanisms. Is strategic thinking innate (or developed at a very young age)? Acquired gradually? The result of experience or repeated exposure? Impacted by the environment in which we grow?

The experimental literature on developmental decision-making has devoted significantly more attention to behavior in individual choice paradigms compared to games of strategy.[3] Existing studies (Sher et al., 2014; Czermak et al., 2016; Chen et al., 2016; Brocas et al., 2017; Fe et al., 2020) point to behavioral age differences but leave several key questions unanswered. First, evidence is often reported on a snapshot of the developmental trajectory. Second, paradigms are often complex and require the aggregation of many abilities to achieve rational play, thereby introducing a confound between analytical ability, forward looking behavior, beliefs about others' choices and payoff maximizing considerations. Developmental psychologists have also studied cognitive sophistication using the Recursive Thinking (Miller et al., 1970; Eliot et al., 1979; van den Bos et al., 2016) and Theory-of-Mind paradigms (Perner and Wimmer, 1985; Wellman et al., 2001), two abilities required to play at equilibrium in strategic settings. However, this literature does not tell if, when and in what contexts these abilities are transferred to strategic play. The present article directly tests the developmental trajectory from childhood to adulthood of strategic sophistication in strict-dominance solvable games.

Dominance solvable games are particularly appealing because steps of reasoning offer a natural algorithm to solve them. Sophistication is defined naturally as the number of steps of reasoning that a subject is able to implement to get closer to the Nash equilibrium. Our objective is to investigate the relationship between that form of sophistication and

[1] See e.g., Costa-Gomes et al. (2001); Johnson et al. (2002); Costa-Gomes and Crawford (2006); Brañas-Garza et al. (2011); Arad and Rubinstein (2012); Brocas et al. (2014, 2018); Kneeland (2015) out of a long list. See also the survey by Crawford et al. (2013).

[2] See e.g., Brañas-Garza et al. (2012); Gill and Prowse (2016); Proto et al. (Forthcoming).

[3] See for example the surveys by Sutter et al. (2019) and List et al. (2018).

age. Notice that we are not interested in situations where more steps of reasoning do not move the individual closer to Nash equilibrium (as for example in the 11-20 money request game of Arad and Rubinstein (2012)) and/or do not result in higher empirical payoffs (as for example in the traveler's dilemma game of Capra et al. (1999)). Instead, we want a setting in which (i) steps of reasoning provide the algorithm to play Nash, (ii) subjects with higher levels of reasoning invariably play closer to Nash and (iii) they obtain higher payoffs. This allows us to rank unambiguously the sophistication of participants.[4]

For this, we need to design a paradigm such that the ability to recursively think about others' behavior simultaneously facilitates the formulation of the (theoretical) equilibrium and the (empirical) payoff-maximizing strategy. We confront two major challenges for the age evolution analysis to be feasible. First, the problem needs to be sufficiently simple and transparent that young children can understand it (without being trivial for high schoolers and young adults). This consideration precludes the use of some standard paradigms, such as the two-person p-beauty contest (Costa-Gomes and Crawford, 2006). This type of games is intuitive for game theorists, replete with interesting and testable properties, but excessively intangible for the minds of children. Second and related, it is critical to minimize the abstract and formal structure of the game. These aspects may lead the intrinsic logical ability necessary for strategic thinking to confound with mathematical or analytical skills (a competence that is expected to develop during adolescence and facilitated by extra years of schooling). They make standard normal-form representations (Costa-Gomes et al., 2001; Kneeland, 2015; Brocas et al., 2018) also unsuitable for our study.[5]

To minimize these concerns, we propose a novel graphical interface where subjects possess three objects with three attributes each: a shape, a color and a letter. Their goal is to select an object with a certain characteristic, which depends on the object selected by another player in the game. This is true for all but one player, who must simply match a feature of a specific single object. This player's decision constitutes the starting point of the iteration process, and the problem of the other players can be iteratively solved by successive elimination, with a maximum of three steps of reasoning.

To analyze the developmental trajectory of behavior in our paradigm, we recruited three populations. The first experiment involves a population of children and adolescents (8 to 18 years old) recruited at a single private school in Los Angeles as well as a control young adult population from USC. This experiment tests the effect of age on strategic

[4] Fe et al. (2020) provides an interesting study of a simplified 11-20 money request game with 5 to 12 years old children.

[5] The difficulty to understand a game when it is presented abstractly has been recognized (Chou et al., 2009; Cason and Plott, 2014). We believe it is exacerbated in the case of children.

sophistication. In the second experiment, we recruited younger children (5 to 8 years old) from that same school and we implemented a simpler version of the same game. This experiment is designed to assess whether the skills detected in children older than 8 years old are already developing before that age. Last, we recruited a third population of middle schoolers (11 to 14 years old) from a single public school also in Los Angeles. This experiment aims to inform us on the potential impact of school characteristics and student demographics on sophistication.

Our analysis yields three main results. First, the vast majority of participants either play always at equilibrium or they play at equilibrium only when they have a dominant strategy. There are few random players, and virtually no one exhibits an "intermediate" level of reasoning (i.e., plays at equilibrium when it requires two steps of reasoning but not when it requires three steps). This is in sharp contrast with the existing adult literature that emphasizes large heterogeneity in levels of reasoning and abundance of intermediate types (Costa-Gomes et al., 2001; Johnson et al., 2002; Costa-Gomes and Crawford, 2006; Brañas-Garza et al., 2011; Brocas et al., 2014; Kneeland, 2015; Gill and Prowse, 2016; Brocas et al., 2018). Second, although age is an important determinant of equilibrium thinking, there is an ability component that is either innate or acquired at a very young age. Furthermore, the evolution over the entire window of observation is not as steep as one might expect. Indeed, the proportion of individuals who consistently play at equilibrium is significantly above 0 at 8 years old (24%) and significantly below 1 at 17 years old (59%). Third and related, the change in equilibrium play is not constant. Choice improves significantly between 3$^{\text{rd}}$ and 5$^{\text{th}}$ grade and stabilizes afterwards. In other words, the contribution of age to equilibrium behavior vanishes relatively early in life (between 12 and 13 years old). Our data reveals important predictors of performance. We find that female participants and subjects with a self-reported preference for science subjects perform significantly better. Finally, differences across schools and across tracks within schools are also associated with differences in sophistication. In particular, we find that students enrolled in different programs or in different GPA-based tracks within programs exhibit different levels of sophistication. Overall, even though the main pattern of behavior (namely the absence of an intermediate level of reasoning) is replicated in all populations, the distribution of sophistication is modulated by individual and group characteristics.

2 Experiment

2.1 Design and procedures

We study a three-person, simultaneous move, complete information strict dominance-solvable game. Working with a population of children and adolescents presents important methodological challenges.[6] We follow the guidelines proposed in Brocas and Carrillo (2020b) to address those obstacles.

Participants. We report the results of three experiments, referred to as MAIN, KING and YOUNG, with a total of 721 children and adolescents and 60 young adults. They all feature the same paradigm but focus on different populations. Populations are described in detail at the beginning of each analysis section.

- MAIN. Our main population consisted of 234 school-age participants from 3rd to 11th grade, studying at the Lycée International de Los Angeles (LILA), a French-English bilingual private school in Los Angeles. For comparison, we also recruited 60 students at the University of Southern California (USC).

- KING. We ran the same experiment with 370 middle schoolers from Thomas Starr King Middle School, a public school in Los Angeles, located less than a mile away from LILA. We used the school's classification criteria to group participants according to their academic achievements.

- YOUNG. Finally, we ran a simplified version of the experiment with 117 younger children from LILA in grades K, 1st and 2nd.

Tasks. The experiment had two tasks implemented on PC tablets, programmed on 'Multistage Games' and always performed in the same order.[7] The first task consisted of two trials of a "lying game," where subjects privately rolled a dice and were rewarded according to the number they reported. The findings of this project are discussed in a different article (Brocas and Carrillo, 2019). The second task, which is the focus of this article, consisted of 18 trials of a three-person, simultaneous move game.

We designed a simple, graphical interface, which was both accessible and appealing to children as young as 8 years of age. The main challenge was to be able to pinpoint the ability of children to think logically and recursively through the game and to avoid confounding effects from other abilities. In particular, it was of paramount importance that differences in behavior reflected as much as possible developmental differences in *logical abilities* rather than developmental differences in mathematical skills and capacity to concentrate for long periods on abstract instructions. This ruled out payoff matrices and other formal presentations standard in the literature with college undergraduates.[8]

[6] Children have limited attention. They respond differently to incentives. They have a limited ability to grasp abstract representations. They do not develop uniformly. Finally, their behavior is best understood if it can be contrasted with that of adults.

[7] Instructions to download the software can be found at `http://ssel.caltech.edu:8000/multistage`.

[8] See Brocas and Carrillo (2020b) for further discussion on the importance of adapting the presentation and instructions to the population.

We designed a simple paradigm in which subjects were matched in groups of three and assigned a role as player 1, player 2 or player 3, from now on referred to as role 1, role 2 and role 3. Each player in the group had three objects, and each object had three attributes: a shape (square, triangle or circle), a color (red, blue or yellow) and a letter (A, B or C). Players had to simultaneously select one object. Role 1 would obtain points if the object he chose matched a given attribute of the object chosen by role 2. Similarly, role 2 would obtain points if the object he chose matched a given attribute of the object chosen by role 3. Finally, role 3 would obtain points if the object he chose matched a given attribute of an extra object. The attributes to be matched were different for different roles and specified by the experimenter. Accordingly, in each game any number of participants could obtain points. All options and objectives of players were common knowledge and displayed on the computer screen. Figure 1 provides a screenshot of the game as seen by role 2.

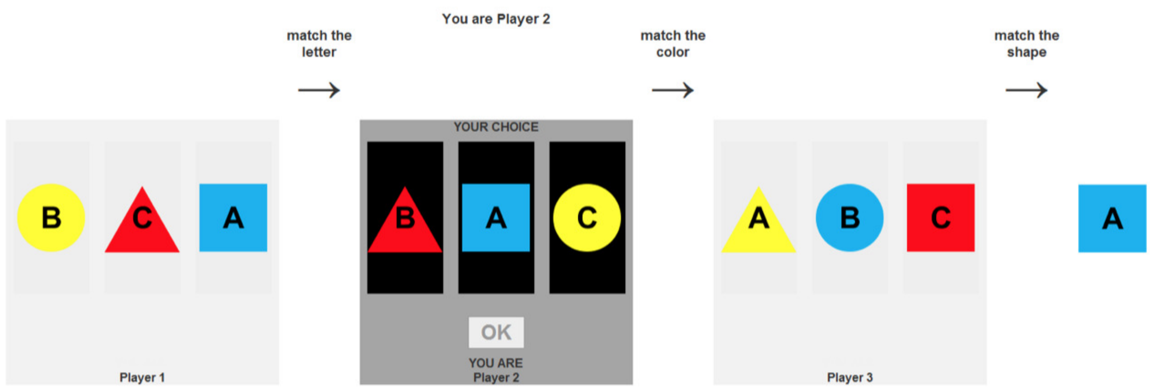


Figure 1: Screenshot of the game (as seen by role 2)

The game can be easily solved with an inductive argument starting from role 3. In the example of Figure 1, role 3 has to match the shape of the outside object, so he obtains the points if he chooses the red square C. Conditional on that choice, role 2 obtains points if he chooses the red triangle B and, again conditional on that choice, role 1 obtains points if he chooses the yellow circle B (the original software uses easily distinguishable colors).[9]

Participants played 6 trials in each role r ($\in \{1, 2, 3\}$) for a total of 18 trials, with anonymous partners randomly drawn after each trial and without feedback. After each trial, participants changed roles and the software changed the shapes, colors, letters and attributes to be matched. To ensure comprehension, we implemented a quiz before the 18

[9]Obviously, if role 2 (3) does not play the equilibrium strategy, then role 1 (2) does not obtain points by playing at equilibrium.

incentivized trials.[10]

We used 9 combinations of objects for the first 9 trials and we repeated them for the following 9 trials. This allowed us to study learning by comparing the choices in the first and second half of the experiment. Also, to separate as much as possible between equilibrium reasoning and chance, we deliberately introduced focal, non-equilibrium objects in roles 1 and 2 of all 18 trials.[11] A transcript of the instructions and quiz is included in Appendix B.

Payoffs. Subjects accumulated points. Following Brocas and Carrillo (2020b), we implemented three different conversions depending on the population. LILA students from grade 6th and above, KING students and USC subjects had points converted into money paid immediately at the end of the experiment in cash (USC) or with an amazon e-giftcard (LILA and KING, where cash transfers on premises are not allowed). USC subjects accumulated \$0.50 and \$0.20 per successful and unsuccessful trial respectively, with a \$5 show-up fee. LILA and KING subjects accumulated \$0.40 and \$0.20 per successful and unsuccessful trial with no show-up fee.[12] The entire experiment lasted less than one school period (between 40 and 50 minutes). Average earnings (on the entire experiment and not including show-up fees) were \$15.2 (USC), \$11.2 (LILA) and \$10.8 (KING).

For elementary school subjects at LILA (grades K to 5th), we set up a shop with 20 to 25 pre-screened, age-appropriate toys and stationery (bracelets, erasers, figurines, diecast cars, trading cards, apps, calculators, earbuds, fidget spinners, etc.). Participants accumulated 40 points and 20 points per successful and unsuccessful trial respectively, and each toy had a different point price. Before the experiment, children were taken to the shop and showed the toys they were playing for. They were instructed about the price of each toy and, for the youngest subjects, we explicitly stated that more points would result in more toys. At the end of the experiment, subjects learned their point earnings and were accompanied to the shop to exchange points for toys. We made sure that every child earned enough points to obtain at least three toys. At the same time, points were always valuable and no child ended up with more points than toys they liked.[13]

[10] Subjects had to answer 4 questions. If they missed one or more, a warning sign would appear stating "not all answers are correct, please try again". The experiment started only after all subjects in a session had completed the quiz correctly.

[11] In the example of Figure 1, the blue square A was the extra object to the right of the screen, and it was also in the choice set of roles 1 and 2. However, no player should, in equilibrium, select it. Appendix A3 reports heuristic rules based on focal objects.

[12] Incentives were calibrated to account for differences in marginal value of money and opportunity cost of time. We provided a positive payment for unsuccessful trials to artificially reduce variance and ensure a pleasant experience of our school-age participants.

[13] The procedure emphasizes the importance of accumulating points while making the experience enjoyable. At this age, a toy is also a significantly more attractive reward than money. Most children are

Questionnaire. We collected demographic information consisting of "gender", "age", "grade", "number of siblings" and "favorite subject at school".

2.2 Theory and hypotheses

Consistent with the experimental literature on dominance solvable games reviewed in the introduction, we expect that participants would differ in their ability to iteratively eliminate dominated strategies. More precisely, we anticipate to find four types of individuals: R (subjects who always play randomly), D_0 (subjects who play at equilibrium only if they have a dominant strategy), D_1 (subjects who play at equilibrium when they have a dominant strategy and can best respond to a D_0 type), and D_2 (subjects who can play as D_0 and D_1, as well as best respond to D_1). These types map well into *nested* levels of strategic sophistication from lowest (R) to highest (D_2). The game, however, does not allow us to distinguish between levels D_2 and above (D_3, Nash, etc.). Notice that in our setting, as in other dominance-solvable games (Kneeland, 2015; Brocas et al., 2018), there is a one-to-one correspondence between steps of dominance and level k theory under random behavior for level 0 and uniform error distribution for levels 1 and above (L_k coincides with D_{k-1} for all $k \geq 1$ and L_0 coincides with R).

The predicted behavior is simple. R plays the equilibrium strategy 1/3 of the time in all roles; D_0 always plays the equilibrium strategy in role 3 and 1/3 of the time in roles 1 and 2; D_1 always plays the equilibrium strategy in roles 2 and 3 and 1/3 of the time in role 1; and D_2 always plays the equilibrium strategy. An immediate implication of this behavioral theory is that we should never observe a subject playing the equilibrium strategy significantly more often in an earlier role than in a later role.

We formulate the following hypotheses.

Hypothesis 1 *The behavior of the vast majority of individuals at all ages is consistent with one of the four types: R, D_0, D_1 or D_2.*

Hypothesis 1 states that the behavioral model that has proved successful to describe the reasoning process of adults in dominance-solvable games (steps of dominance or level k) is expected to fit also the behavior of children and adolescents.

Hypothesis 2 *There are few or no D_2 types in our youngest school-age subjects. There are no R and few or no D_0 types in our oldest school-age and adult subjects. There are D_1 types in all ages.*

familiar with this method of accumulating points or tickets that are subsequently exchanged for rewards since it is commonly employed in arcade rooms and fairs. We spent an average of \$4 in toys per child.

Informally, the idea behind Hypothesis 2 is that the game is hard to solve for young children, very easy to solve for young adults, and subject to significant improvements with age. Also, and in line with the existing literature on adult behavior (e.g., Costa-Gomes and Crawford (2006); Brocas et al. (2018)), a significant fraction of subjects is expected to perform a positive but limited number of steps of reasoning (D_1 or, equivalently, L_2 under the level k interpretation).

Hypothesis 3 *There is a gradual and strictly monotonic shift in types with age, from lowest to highest level of sophistication (R to D_0 to D_1 to D_2).*

According to Hypothesis 3, sophistication increases steadily with age. We also anticipate higher sophistication in adults than in our oldest school-age students.

While we think that Hypotheses 1, 2 and 3 are natural, they carry important implications for developmental decision-making. Indeed, validating them would show that the behavioral theory that captures heterogeneity in decision making by adults in dominance-solvable games, is also successful in explaining the choice of a young population. In other words, cognitive reasoning is quantitatively different but qualitatively similar across ages, with a smooth improvement in performance with age. It would also imply that, over time, we develop abilities that facilitate performing more steps of reasoning. More specifically, these abilities allow us to progress from random behavior, to one step of reasoning, then two steps and finally three steps.

3 Evolution of behavior with age

We first study the choices by LILA students from 3$^{\text{rd}}$ to 11$^{\text{th}}$ grade, and compare them to the control USC undergraduate population (U). Students at LILA are 73% White, 16% mixed races, 5% Hispanic and 4% Asian. It is a homogenous population (same school, same curriculum, similar social and economic backgrounds), which is helpful for comparisons across grades. Table 1 summarizes the number of subjects by grade.

Population	LILA									USC
Grade	3	4	5	6	7	8	9	10	11	U
# subjects	34	17	20	41	27	18	25	23	29	60

Table 1: Summary of participants

While the number of students per grade is relatively small, participation rates are high (61% of students in those grades took part in the study). We ran 20 sessions with school-age students in a classroom at LILA and 5 sessions with undergraduates at the

Los Angeles Behavioral Economics Laboratory (LABEL) at USC. Sessions had 9, 12 or 15 participants and followed *identical* procedures. For each school-age session, we tried to include subjects from the same grade, but for logistical reasons we sometimes had to mix subjects from two consecutive grades.

Notice that most studies with children do not recruit an adult population. We believe it is important to include an adult control group to establish a behavioral benchmark, even if the comparison is imperfect (Brocas and Carrillo, 2020b). In our case, the majority of students at LILA are from upper-middle socioeconomic status. After graduation, they typically attend well-ranked colleges (including USC). Overall, while there are some important differences in terms of nationality (higher fractions of europeans at LILA), ethnicity (higher fraction of asians at USC), and size of peer group (larger and more anonymous cohorts at USC) among other characteristics, the two populations match reasonably well.

3.1 Aggregate choice

Figure 2 reports the average number of equilibrium choices by grade and role.

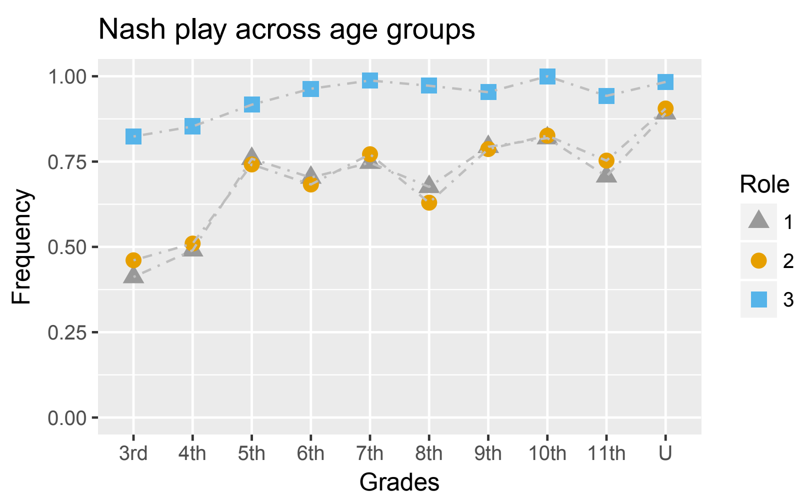


Figure 2: Proportion of equilibrium choices by grade and role

Within grades, equilibrium play is significantly higher in role 3 than in the other roles for all grades. By contrast, and to our surprise, equilibrium play is not significantly different between roles 1 and 2 for any grade. The behavior across grades also unveils interesting patterns. Equilibrium behavior in roles 1 and 2 is similar in 3$^{\text{rd}}$ and 4$^{\text{th}}$ grade, increases significantly in 5$^{\text{th}}$ grade, and remains constant afterwards (there is a dip in 8$^{\text{th}}$ grade, though it is not statistically significant). Compared to middle school, the USC population plays Nash only marginally more often.[14] Finally, in all grades and roles the probability of equilibrium behavior is above 0.33. Therefore, the best response to the empirical behavior is to play the equilibrium strategy for all roles and grades. In other words, deviations from Nash cannot be explained by non-equilibrium behavior as a best response to the empirical strategy of others.

3.2 Individual analysis

Although aggregate behavior is instructive, patterns of choice at the individual level are more revealing. According to section 3.1, playing the equilibrium action is also the payoff-maximizing, best response strategy to the empirical behavior of the population. Table 2 reports the fraction of individuals who play the equilibrium action in all rounds (18) or makes at most one mistake (17-18).

	LILA									USC
	3	4	5	6	7	8	9	10	11	U
18	0.21	0.12	0.30	0.37	0.44	0.39	0.40	0.78	0.52	0.69
17-18	0.24	0.12	0.60	0.46	0.59	0.39	0.64	0.78	0.59	0.85

Table 2: Fraction of equilibrium players by grade

Non-equilibrium players are frequent except for 10$^{\text{th}}$ graders and USC students. The patterns confirm also the aggregate analysis: we observe low levels of equilibrium compliance in 3$^{\text{rd}}$ and 4$^{\text{th}}$ grade, an increase in 5$^{\text{th}}$ grade and a stabilization thereafter (with a statistically not significant dip in 6$^{\text{th}}$ and 8$^{\text{th}}$ grades and a statistically significant peak in 10$^{\text{th}}$ grade and USC).

Given the significant proportion of individuals who do not play the equilibrium strategy, we next classify subjects into types. We use a very simple method. We label the behavior in role r "equilibrium" if the subject played the equilibrium action 5 or 6 times (out of 6) and "random" if the subject played the equilibrium action 0, 1 or 2 times (out of 6).[15] We then use the theory developed in section 2.2 to classify individuals into R, D_0, D_1 and D_2. The remaining subjects are classified as O, for "Other".[16] Note, however,

[14] We adjusted for multiple hypothesis testing via the Holm's method (that control for the Family Wise Error Rate) and the less stringent Benjamini and Hochberg's False Discovery Rate (FRD) method (that controls the proportion of false positives among the set of rejected hypotheses). Unless otherwise stated, the results reported here are robust to both corrections.

[15] The measure is moderately conservative as it allows one mistake but remains agnostic on individuals who play the equilibrium action 3 or 4 times.

[16] Alternatively, we could structurally estimate types using maximum likelihood methods. This is superior only when data abounds and when subjects do not fall neatly into types. Given our data, our simple classification is more revealing.

that this method could potentially leave unclassified a significant number of individuals (those who play 3 or 4 times the equilibrium action in any role, and those who play more often the equilibrium action in an earlier than in a later role). Table 3 summarizes our classification method.

	role 1	role 2	role 3
R	random	random	random
D_0	random	random	equilibrium
D_1	random	equilibrium	equilibrium
D_2	equilibrium	equilibrium	equilibrium

random = 0-1-2 (out of 6); equilibrium = 5-6 (out of 6)

Table 3: Criterion for classification of subjects into types

Figure 3 reports the proportion of subjects by grade who are classified under each type, from most sophisticated (bottom) to least sophisticated (top).

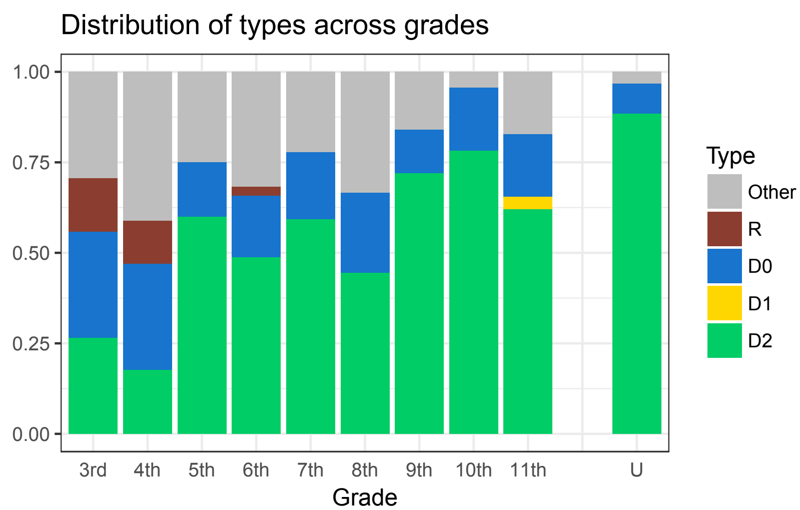


Figure 3: Proportion of subjects by type and grade

In strong support of Hypothesis 1, our theoretical model provides a very solid behavioral template. Indeed, the choice of 76% of LILA students and 97% of USC students can be accounted for by one of the four types described in section 2.2. The proportion of subjects who do not fit in one of these types (O) decreases with age, although it is statistically smaller only for 10^{th} graders. In other words, the level k behavioral theory that has proved successful in explaining non-equilibrium behavior of adults performs well also with children and adolescents.

Result 1 *Hypothesis 1 is supported by the data. The majority of subjects behave consistently with one of the four types of our behavioral model* (R, D_0, D_1, D_2).

At the same time, there are very few R individuals and only one D_1 subject in the entire sample, that is, all the classified individuals are either D_0 (players who can only recognize a dominant strategy) or D_2 (equilibrium players). This is consistent with the result in Figure 2, which highlighted that aggregate equilibrium performance within a grade is very similar in roles 1 and 2. It is in sharp contrast with our Hypothesis 2. Indeed, we expected that subjects would learn gradually with age to perform more and more steps of reasoning. Instead, they either recognize only a dominant strategy or all the steps of reasoning. It is also radically different from the experimental literature that emphasizes large heterogeneity in steps of reasoning in adults. Admittedly, and for the purpose of being accessible to young children, our setting is simpler than most existing games. However, it is also devoid of an analytical framework. We conjecture that part of the reason why some individuals perform some but not all the steps of reasoning in traditional dominance-solvable games is because of the complexity of the formal presentation. In other words, it is possible (and worthy of further investigation) that some intermediate levels of reasoning are the result of limitations in the ability to understand finer aspects of the game and/or mechanically compute all the required steps, as opposed to a limitation in the ability to perform s steps of reasoning after having successfully performed $s-1$.[17]

Also against our Hypothesis 2, 26% of our 3$^{\text{rd}}$ graders are classified as D_2 (21% play the equilibrium strategy in all 18 trials). Conversely, 34% of 11$^{\text{th}}$ graders are classified as D_0 or O (21% play less than 4 out of 12 times the equilibrium strategy in roles 1 and 2). These two population are significantly different from 0.

Result 2 *Hypothesis 2 is not supported by the data. Subjects either recognize only a dominant strategy or always play at equilibrium. Also, some very young players display an innate ability to play always at equilibrium while some young adults are unable to perform two-steps of dominance.*

Finally, and in partial support of Hypothesis 3, we notice a weakly monotonic (but not gradual) increase in strategic types with age. Participants in 3$^{\text{rd}}$ and 4$^{\text{th}}$ grade are mainly type R and D_0 whereas participants in 5$^{\text{th}}$ grade and above are predominantly

[17] On the other hand, it cannot explain all the difference. Indeed, as Costa-Gomes and Crawford (2006) show in an adult population, intermediate types survive when subjects demonstrably understand the structure of the game.

D_2, with only small differences after 5th grade. However, this classification of proportions by grade is not the most adequate for statistical tests. Therefore, at this stage we refrain from making definitive assertions on the evolution of equilibrium behavior with age. A more in-depth study of this question is performed in the regression analysis of section 3.3.

Taken together, our findings suggest that the ability to solve dominance-solvable games develops differentially. While this ability is acquired instinctively by some young children, it eludes some educated young adults.[18] It also seems that the developmental trajectory plateaus (or at least decelerates) at a relatively young age, around 5th grade.

Finally, it is important to clarify that our design cannot distinguish between a subject who does not perform s steps of reasoning and a subject who does but believes that the subject they are paired with does not perform $s - 1$ steps. This distinction is at the core of the work by Kneeland (2015) who proposes "ring games" capable of disentangling orders of rationality.[19] In the paper, we have associated deviations from equilibrium with self-limitations. While such interpretation could be challenged (and it would be interesting to propose new paradigms to further investigate that distinction), there are some indications that it is a reasonable one. First, it is implausible (and severely incorrect from an empirical viewpoint) that subject of any age in role 2 would believe that their role 3 counterpart will not play the equilibrium action with high probability. Second and more importantly, if non-equilibrium choices were due to beliefs about rationality of others, we would presumably observe more deviations in role 1 (where an equilibrium choice requires third-order rationality in Kneeland (2015)'s terminology) than in role 2 (where an equilibrium choice requires only second-order rationality). Instead, no such D_1 types are present in our pool.

3.3 Regression analysis

To better understand the determinants of equilibrium behavior, we perform a series of OLS regressions at the individual level. We only consider the 234 school-age students, to avoid biasing the results with the undergraduate population. In columns 1, 2 and 3 of Table 4, the dependent variable is the percentage of equilibrium choices of each participant in role r $(\in \{1, 2, 3\})$. Our main independent variable is the *Age* in months of the participant at the date of the experiment. We include a dummy variable for favorite topic at school ($STEM = 1$) to account for analytical inclination. STEM refers to a reported preferences for Mathematics, Sciences or Technology. Consistent with the curriculum of the school,

[18] The fact that most high schoolers and young adults who do not play consistently at equilibrium are classified as D_0 and not R suggests that they have paid attention to the game.

[19] This work has subsequently been extended by Friedenberg et al. (2018) to distinguish between rationality bounds and strategic bounds.

the other categories offered were English, French, History/Geography and Arts/Music, which we globally refer to as 'Arts & Humanities'. We also add demographic dummy variables for gender (*Male* = 1) and whether the participant has siblings (*Siblings* = 1).

	OLS			Spline OLS		
	Role 1	Role 2	Role 3	Role 1	Role 2	Role 3
Age	0.003***	0.003***	0.001***	—	—	—
	(0.001)	(0.001)	(0.0003)			
Age (spline1)	—	—	—	0.009***	0.007***	0.004***
				(0.002)	(0.002)	(0.001)
Age (spline2)	—	—	—	-0.001	-0.0001	-0.001
				(0.001)	(0.001)	(0.001)
knot (months)	—	—	—	147	148	154
STEM	0.191***	0.224***	0.037	0.196***	0.228***	0.039
	(0.049)	(0.049)	(0.024)	(0.047)	(0.048)	(0.024)
Male	-0.082*	-0.104**	-0.010	-0.086*	-0.108**	-0.014
	(0.047)	(0.048)	(0.023)	(0.046)	(0.047)	(0.023)
Siblings	-0.059	-0.051	-0.019	-0.069	-0.058	-0.022
	(0.053)	(0.054)	(0.027)	(0.052)	(0.053)	(0.026)
Constant	0.240**	0.261**	0.744***	-0.533**	-0.319	0.453***
	(0.121)	(0.123)	(0.060)	(0.245)	(0.246)	(0.107)
Observations	234	234	234	234	234	234
Adj. R^2	0.112	0.123	0.048	0.156	0.147	0.086

(standard errors in parenthesis) * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 4: Regressions of equilibrium choice by role

As expected, *Age* is a key determinant of equilibrium behavior. Males play at equilibrium less often than females in roles 1 and 2. There is also a strong explanatory power of the self-reported preferred school topic. Indeed, participants who report a preference for STEM play the equilibrium action 20% more often in roles 1 and 2 than those who prefer Arts & Humanities. These effects may reflect the philosophy of the school. Indeed, LILA offers extra-curricular activities around STEM topics (math olympiads, math kangaroo, STEM club, robotics) and encourages students to develop their scientific skills. It is plausible that students who take advantage of these opportunities develop faster and more consistently their logical thinking. Also, the school promotes female confidence by making STEM projects attractive to them,[20] inviting keynote female speakers to discuss

[20]STEM projects often revolve around robotics and tend to attract mostly males. To correct this bias, the school has added activities that may be of special interest for females, such as architectural, urban and fashion design projects.

their career choices and achievements, and encouraging advocacy against gender-based discrimination.[21]

The analysis in sections 3.1 and 3.2 suggests an increase in equilibrium behavior with age but also a deceleration after a certain grade. To further investigate these dynamic trends, we conduct a spline regression analysis to estimate the age at which such deceleration occurs for each role. The method consists of running OLS regressions assuming that a kink exists, and in identifying the kink that provides the best R^2-based fit. We include the same controls as above. The spline regressions by role are reported in columns 4, 5 and 6 of Table 4 and represented in Figure 4.

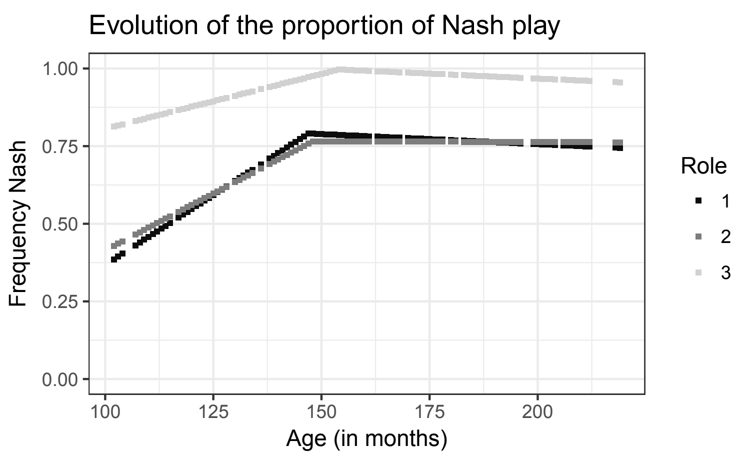


Figure 4: Spline OLS regression

The regressions strongly support our previous conclusions. Performance in roles 1 and 2 increases significantly up to a certain age (around 12 years old), and then stabilizes approximately at 0.75. Fits are virtually identical for both roles; the slopes are not statistically different and the knots are estimated to be one month apart. The same trend is present in role 3, although the performance increase before 7$^{\text{th}}$ grade is significantly smaller than in roles 1 and 2. It suggests that other cognitive skills–such as attention and task switching–which are known to develop during elementary school (Chelune and Baer, 1986), may contribute to performance in the simplest choice problem.

Result 3 *Hypothesis 3 is weakly supported by the data. Equilibrium performance increases with age very significantly during elementary school but it stabilizes in 6th grade.*

[21]For instance, middle and high school students organize events in conjunction with Girl Up advocacy group, founded by the United Nations Foundation.

3.4 Other analyses

The main analysis suggests that age and a preference for STEM are strong predictors of Nash play at the aggregate level. We show in Appendix A1 that these variables also predict differences in individual performance and type. In Appendix A2, we split the sample between students who report a preference for STEM and those who prefer Arts & Humanities, to isolate the differences in the developmental trajectory of these two groups. For all grades above 5$^{\text{th}}$, students with a STEM preference play Nash more than 75% of the time in roles 1 and 2, while their counterpart never reach that performance, except in 10$^{\text{th}}$ grade.

The similarity between behavior in roles 1 and 2 also points to a common reasoning process, one that is unlikely to involve beliefs about others. We investigate in Appendix A3 whether non-equilibrium players follow a discernible strategy. We show that, when playing in roles 1 and 2, D_0-types consistently pick the focal, non-equilibrium object (as mentioned in section 2.1). In the absence of a dominant strategy, these subjects are misguided into choosing a suboptimal strategy. This pattern is less pronounced for types R or O.

In Appendix A4, we report a moderate but statistically significant increase in equilibrium behavior between the first and second half of trials, especially in role 1. Some subjects evolve to a more sophisticated type over time (from D_0 to D_2). This suggests that, despite the absence of feedback, playing in a certain position (e.g., role 3) helps a small but positive fraction of subjects better understand how to play in other positions (e.g., role 2 then role 1).

4 Strategic thinking in middle school

Middle school is an age specially important in our paradigm for several reasons. First, it is a key transitional period from childhood to young adulthood. From a physiological viewpoint, adolescence is a critical time for the development of the brain's neural network. Changes occurring during middle-school have a crucial impact on cognitive and emotional responses (Choudhury et al., 2006). From and educational viewpoint, subjects move from small classrooms with one or two teachers and close academic supervision to larger classrooms, different teachers for every subject and an expectation of academic maturity, responsibility and independence (in organizing schedules, completing homework, etc.). Second, existing research using indirect (Harbaugh et al., 2001) as well as direct (Brocas et al., 2019) tests of transitivity show that by the age of 10 (but not earlier), individuals are as rational decision makers as adults. Third and related, according to the results in section 3, by the beginning of middle school equilibrium behavior in our game is at steady state, with no significant age-related improvements afterwards.

By conducting the same experiment on a different population of adolescents, we can address two questions. First, are logical abilities necessary for level k reasoning dependent on the environment in which learning takes place? In particular, are there behavioral differences across schools, curricula and socio-economic environments? Second, can we replicate the flat developmental trajectory observed at LILA in a different population of middle schoolers? Our second experiment conducted at Thomas Starr King Middle School ("KING"), a large public school with three magnets and several academic tracks, provides a unique chance to compare the behavior of 11 to 14 years old individuals in strategic games as a function of the economic and educational background (LILA v. KING) as well as the academic characteristics (magnets and tracks within KING).

4.1 Preliminaries

The backgrounds in KING and LILA are very different, even though the schools are located less than one mile apart. The two schools differ in curriculum (bilingual in LILA v. monolingual in KING), class size (less than 20 students per class at LILA compared to 35 at KING, except in special education classes), school size (around 200 middle schoolers at LILA and 2000 at KING) and peer group (many students at LILA remain together from preK to 12th grade whereas KING comprises only middle schoolers coming from different elementary schools in the LA area).

KING offers three sharply differentiated magnets with a focus on visual arts, environment and technology/arts, respectively: Film and Media (FILM), Environmental STEAM (ENV) and Gifted/High Ability in Technology and Arts (GIFT).[22] The majority of students in FILM are Latino (55%) followed by White (20%) and Asian (12%). Most students are of low socioeconomic status (75% live at or below the national poverty level). Only a minority of students end up going to college (typically the local community college). The ethnic composition of ENV is similar (55% Latino, 28% White and 9% Asian) but no student lives below the national poverty level. Students attending GIFT have a similar socioeconomic status as ENV (0% is living below the poverty level) but a slightly different ethnic composition (33% Latino, 30% White and 24% Asian).

Academic differentiation by topic and talent is a core component of public education in the Los Angeles Unified School District (LAUSD, 2020), and KING is a paradigmatic case. Compared to ENV and FM, only children satisfying specific eligibility criteria and identified as academically advanced (gifted) can opt in the GIFT magnet.[23] Magnets fur-

[22] For more information, see `https://www.kingms.org`.

[23] Students need to be identified as 'gifted' by a psychologist from Los Angeles United School District in the Intellectual, High Achievement, Specific Academic Ability, Creative Ability, or Leadership Ability categories.

ther separate students into three tracks: challenged, regular and honors. The "challenged" track is a mix of children with mild learning disabilities (dyslexia, problems focusing, etc.) and special needs (english learners), although the majority are children at academic risk (low attendance, low GPA). The "honors" students are children with a higher GPA than their peers. A student could be in the honor class for one topic and the regular class for a different one, so we used the classification in the class of the teacher who granted us access to the students. Naturally, no challenged track is offered in the GIFT magnet.

We conducted the experiment with 370 middle schoolers at KING. Table 5 summarizes the participants by grade, track and magnet.

		Challenged	Regular	Honors
FILM	6^{th}	21	0	0
ENV	6^{th}	0	76	100
GIFT	6^{th}	0	56	28
FILM	7^{th}	46	0	0
FILM	8^{th}	43	0	0

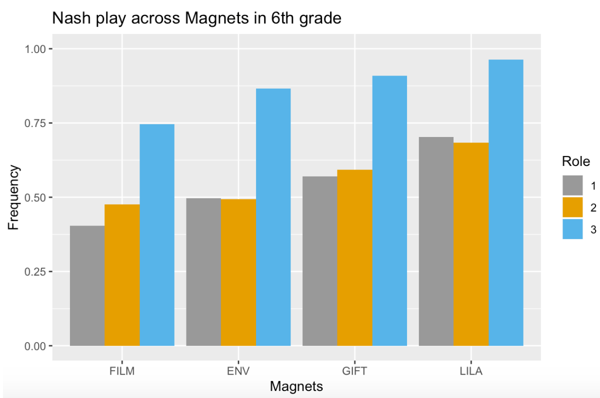
Table 5: Participants at KING by grade, track and magnet

Two clarifications are in order. First, only a few teachers (those with classroom availability and no conflict of schedule or testing) granted us access to their class, which explains the large number of entries with zero students. However, for the classes where participation was an option, we obtained consent from a vast majority of children (89%). Second, we employed the exact same protocol as with LILA students from 6^{th} to 11^{th} grade, including classroom layout, interface, instructions, payment method and conversion rate.

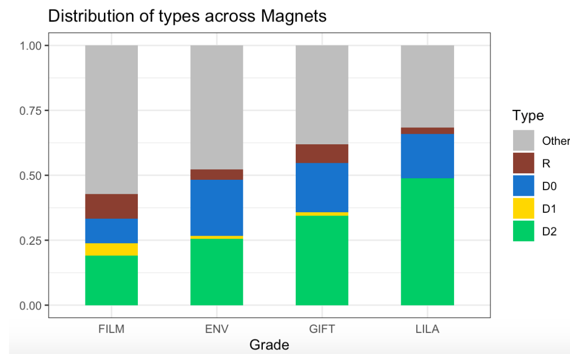
Given the available populations, we conduct two studies separately. First, we analyze the behavior in the three magnets of 6^{th} graders at KING (FILM, ENV and GIFT). We preferred a classification by magnet rather than track because, within our sample, magnet is a better indicator of academic ability (all challenged are in FILM, all gifted students are in GIFT, and separation between regular and honors is valid only within a subject). We then compare these students to the LILA 6^{th} graders of our previous sample (section 4.2). Second, we compare the evolution through middle school (6^{th}, 7^{th}, 8^{th} grade) of challenged students from the FILM magnet at KING with the LILA students of the same age (section 4.3). We should emphasize that comparisons across magnets and schools should be taken with a grain of salt given the large documented demographic and socioeconomic differences. At the same time, the comparison is useful to identify factors associated with sophistication.

4.2 A comparative analysis of 6th graders

Figures 5a and 5b report the proportion of equilibrium choice by magnet and role, and the distribution of types across magnets, including a benchmark comparison of LILA subjects.



(a) Equilibrium choices by magnet



(b) Types by magnet

Figure 5: Equilibrium behavior and type classification of 6th graders by track

As we can see from Figure 5a, one major result of the previous section–the similar performance in roles 1 and 2 and the statistically higher performance in role 3–is very robust, as it holds in every magnet. Comparing across magnets in KING, we obtain the expected ranking, with gifted students (GIFT) performing best and challenged students (FILM) performing worst. However, differences are small. Indeed, the difference between GIFT and ENV is marginally significant in role 2 (p = 0.06, False Discover Rate (FDR) adjusted) but not in roles 1 and 3, whereas the difference between ENV and FILM is significant only in role 3 (p = 0.017, FDR adjusted). All tracks perform worse than students at LILA, although the difference between LILA and GIFT is only marginally significant in role 1 (p = 0.078, FDR adjusted).

Figure 5b confirms these results. There are no statistical differences in types across magnets at KING. By contrast, and despite the low number of observations, the proportion of D_2 is significantly higher in LILA than in ENV (p = 0.038, FRD adjusted) but not significantly higher in LILA than in GIFT. Perfect Nash players are also more prevalent in LILA (37%) than in the ENV (11%) or FILM (4%) magnets at KING ($p < 0.04$, FDR adjusted) but not statistically different from GIFT (21%).

Overall, 6th graders in the FILM and ENV magnets at KING behave more like 3rd or 4th graders at LILA, with only 50% of equilibrium behavior in roles 1 and 2 and a large fraction of unclassified subjects. GIFT participants outperform their peers but they lag compared to the children in the other school.

We finally conduct a similar OLS regression as in columns 1, 2 and 3 of Table 4. We remove the *Age* variable (since all the subjects are 6^{th} graders) and include dummies for KING magnets as well as LILA, with FILM being the omitted variable. The results are summarized in Table 6.

	Role 1	Role 2	Role 3
ENV	0.094	0.017	0.119**
	(0.080)	(0.081)	(0.047)
GIFT	0.169**	0.119	0.156***
	(0.086)	(0.086)	(0.050)
LILA	0.302***	0.214**	0.205***
	(0.094)	(0.095)	(0.056)
STEM	-0.003	0.036	-0.045*
	(0.041)	(0.041)	(0.024)
Male	0.017	-0.014	0.005
	(0.040)	(0.040)	(0.024)
Siblings	0.008	-0.013	-0.004
	(0.052)	(0.052)	(0.030)
Constant	0.387***	0.478***	0.770***
	(0.094)	(0.095)	(0.055)
Observations	322	322	322
Adj. R^2	0.029	0.023	0.047

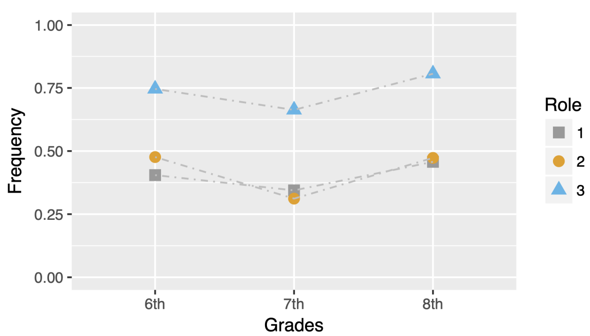
* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 6: Regressions of equilibrium choices

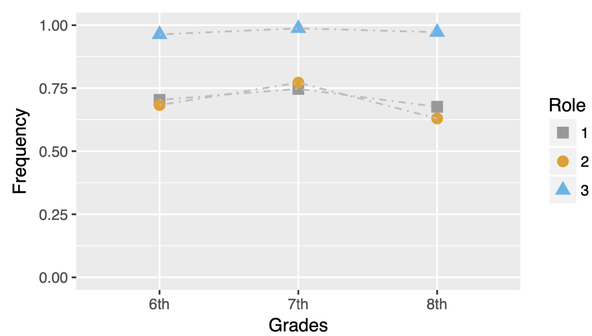
LILA students perform significantly better in all roles and students in GIFT are closer to equilibrium than students in FILM in roles 1 and 3. The difference between ENV and FILM is only significant in the simplest choice problem. In this population, we find no effect of gender or siblings on performance. Because academic inclination is a main component of magnet choice (ENV and GIFT attract students with a penchant for STEAM and technology respectively), the magnet dummy captures STEM ability. This may be the reason why a self-reported preference for STEM, which is very strongly associated with performance at LILA, does not have the same predictive power in the more narrow 6th grade population of children with a large fraction already self-selected to STEM education.

4.3 Evolution during middle school

We next study the evolution during middle school (6^{th} to 8^{th} grade) for the FILM magnet at KING and compare it to the students at LILA with the understanding that, within this magnet, we are only testing challenged students. As mentioned before, 6^{th} grade is a major transition year from an educational viewpoint. We therefore expected that the evolution and adaptation during the three academic years of middle school to a more rigorous and challenging environment would go hand in hand with increased performance in our task. This is not what we obtained at LILA, where we observed a remarkably constant behavior during those years (section 3). However, it is possible that such result was specific to the school. Figure 6 depicts the proportion of equilibrium choices of these two groups of students by role and grade.



(a) Equilibrium: KING (FILM magnet)



(b) Equilibrium: LILA

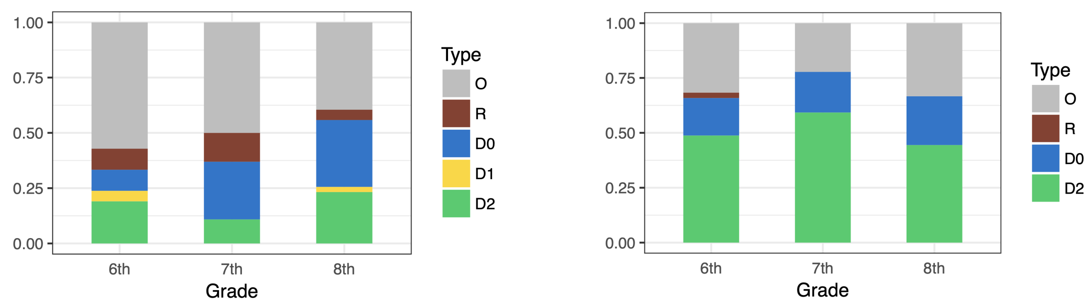
Figure 6: Evolution of equilibrium behavior in middle school (KING and LILA)

Corroborating all previous findings, performance within each school is similar in roles 1 and 2 and higher in role 3 in every grade. Just like in LILA, we also obtain the surprising result in KING that performance does not improve over the middle school years in any role. LILA students perform significantly better in all roles than challenged students in the FILM magnet at KING. In fact, the performance of the latter is rather poor: around 75% of equilibrium choices in the simplest role 3 and only slightly better than random in roles 1 and 2. While the comparison has to be made with extreme caution, according to our data the 8^{th} graders in FILM behave like 3^{rd} or 4^{th} graders in LILA.

Figure 7 compares the types in both populations.

As expected, the fraction of D_2 subjects in KING is small. Also, and contrary to LILA, more than 50% of subjects in 6^{th} and 7^{th} grade cannot be classified in one of our four types. This means not only that many subjects play non-equilibrium actions in the simple role 3, but also that they do not play at equilibrium more often in later than in earlier roles, thereby evidencing significant confusion.

Finally, we run OLS regressions to pinpoint the determinants of equilibrium behavior in each role. The variables we use are age, a dummy for school ($LILA = 1$) and the same controls as previously. The results are presented in Table 7.



(a) Types: KING (FILM magnet) **(b)** Types: LILA

Figure 7: Middle school types in KING and LILA

	Role 1	Role 2	Role 3
Age	0.001	-0.002	0.001
	(0.002)	(0.002)	(0.001)
LILA	0.322***	0.301***	0.259***
	(0.048)	(0.051)	(0.034)
Male	0.027	0.012	-0.062*
	(0.047)	(0.050)	(0.033)
STEM	0.044	0.085	0.102***
	(0.049)	(0.052)	(0.034)
Siblings	-0.003	-0.088	0.038
	(0.064)	(0.068)	(0.045)
Constant	0.283	0.796**	0.491**
	(0.341)	(0.361)	(0.239)
Observations	196	196	196
Adj. R^2	0.175	0.156	0.238

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 7: Regressions of equilibrium choices

Within middle school, age is not a determinant factor of equilibrium choice in any role. LILA subjects perform drastically better than KING subjects in the challenged track. Females and those with a preference for STEM also play closer to equilibrium, but only in role 3. It is worth noting that running the same regression on the KING population alone (omitted for conciseness) yields the same result: equilibrium play is associated with a preference for STEM ($p = 0.018$) and is more prevalent among females ($p = 0.057$) only in role 3. This is interesting because the FILM magnet does not explicitly promote STEM activities and the school does not feature initiatives that promote female confidence. It suggests that LILA activities may only reinforce intrinsic abilities. For instance, it is possible that students who like STEM topics develop those abilities independently, while school offerings permit to enhance these skills.

4.4 Summary

The KING experiment reveals small differences within 6^{th} graders across magnets and larger differences across schools (section 4.2). This casts a warning flag to the practice of pooling data from non-homogenous schools. On the other hand, we do not find any improvement over middle school in any school, suggesting similar qualitative trajectories in this three-year span (section 4.3).

The result indicates that important unobserved environmental factors shape the developmental trajectory of children. The relatively small differences in equilibrium behavior across magnets may reflect a homogeneous learning experience of students at KING, preventing the full translation of academic differentiation into logical play. The significant differences between schools, even when we consider only highly achieving students at KING, should be put in perspective by noting that LILA does not differentiate education according to eligibility criteria and classes mix high and low GPA students. This suggests that differentiation might not be the most efficient tool to promote the development of logical abilities. Differences between schools may also stem from differences in individualized attention due to class size. Finally, even though some students at KING speak two languages, the school curriculum is monolingual. There is converging evidence that bilingual education offers cognitive benefits that translate into decision-making (Bialystok, 2005). Some of the differences we found between schools may also result from this educational feature.

5 Strategic thinking in younger children

Our initial intention was to run the experiment at LILA in all grades, starting from Kindergarten. However, during pilot testing we realized that the game was overwhelming for the youngest participants, due to the amount of information they had to process. Consequently, we developed a simplified version for them. We recruited 38 subjects from K, 37 subjects from 1st and 42 subjects from 2nd grade at LILA for a total of 117 children (a 75% participation rate), which we refer to as the "YOUNG" population. The game consisted of only two players in each group (the analogue of roles 2 and 3), only two attributes (shape and color), and only eight trials (four in each role). Figure 8 presents a screenshot.

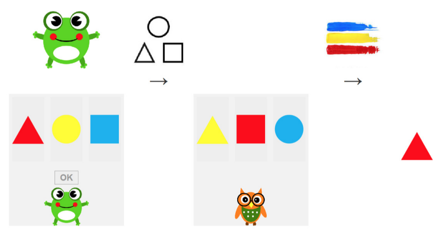


Figure 8: Screenshot for grades K, 1 and 2

The methods were identical: random and anonymous partners in every trial, roles changed after each trial, no feedback between trials, and focal objects (in the example of Figure 8, the red triangle is the extra object and an object in the frog's choice set).[24]

We used a similar procedure as previously to classify subjects. Since we only have the analogue of roles 2 and 3, we cannot distinguish between 2 and 3 steps of reasoning. Table 8 summarizes the classification method for this population (the analogue of Table 3).

	role 2	role 3
R'	random	random
D'_0	random	equilibrium
$D'_{1/2}$	equilibrium	equilibrium

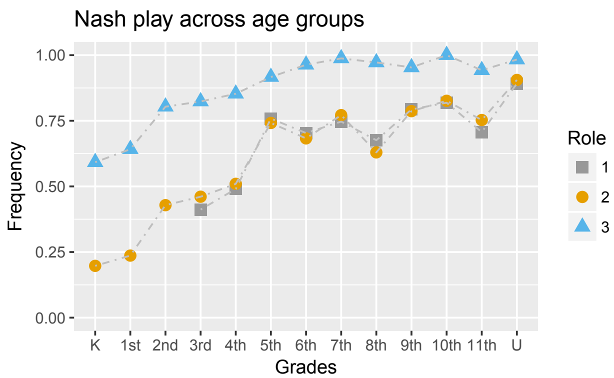
random: 0-1 (of 4); equilibrium: 3-4 (of 4)

Table 8: Classification of YOUNG subjects into types

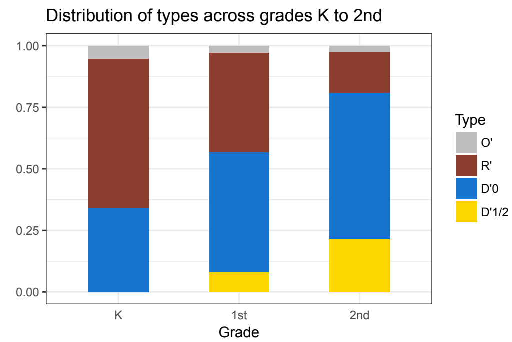
Figure 9a presents the same information as Figure 2, namely the average percentage of equilibrium behavior by grade, except that we add the YOUNG population (roles 2 and 3 for grades K, 1st and 2nd). Figure 9b reports the analogue of Figure 3 (proportion of

[24] Since reading is not an acquired skill at this age, there were no written instructions on the screen: the roles were "frog" and "owl", and the objectives "match the ... " were replaced by graphical descriptions. These presentation changes are minor and introduced only to facilitate comprehension.

types by grade) in the YOUNG population.



(a) Equilibrium choice by grade



(b) Types by grade

Figure 9: Equilibrium behavior and type classification of YOUNG subjects

Equilibrium behavior is not significantly different between K and 1^{st} in roles 2 and 3, and they are both lower than in 2^{nd} ($p < 0.02$, FDR adjusted). Choices between 2^{nd} and 3^{rd} are not statistically different in either role, but remember that the protocol and number of observations are different, so the results are not directly comparable.

Types are also similar in K and 1^{st}: mostly R' and D'_0, with very few equilibrium players (only 1 participant in 1^{st} plays the equilibrium in all 8 trials). By contrast, 9 participants in 2^{nd} grade (21%) are classified as $D'_{1/2}$, of which 6 play the equilibrium in all trials. This group looks similar to 3^{rd} graders (again acknowledging the difficulty to compare these two grades). Overall, the data shows severe difficulties in K and 1^{st} to understand dominant strategies, and no evidence of thinking beyond that. This is consistent with their developing a

There is a leap in understanding in 2^{nd} grade, which seems to stabilize in 3^{rd} and 4^{th}, followed by another increase in 5^{th}. We present some OLS regressions in Appendix A5.

Overall, the YOUNG experiment shows that the developmental trajectory of steps of reasoning starts early, and that sophistication increases with age. While K and 1^{st} are rarely able to play at equilibrium when more than one step of reasoning is necessary, the performance of 2^{nd} graders is similar to that of 3^{rd} graders.

6 Conclusion

In this study, most participants either play at equilibrium or recognize only a dominant strategy. It is very unlikely that rational subjects deviate from equilibrium choices in roles 1 and 2 because they (incorrectly) think that the majority of peers in role 3 cannot solve the game. The similar performance in roles 1 and 2 also suggests that mistakes are not due to pure mathematical limitations or beliefs about rationality of others: if it were the case, we would observe a lower performance in the most challenging role 1. Overall, the data suggest that both under-performance and absence of intermediate levels result from a cognitive limitation in recursive reasoning. Therefore, the ubiquitous intermediate levels reported in previous experiments that test level k theories are likely due to features that are not present in our design. They may stem from concerns regarding the ability of others to reach a decision, from complex designs requiring long iterations towards the equilibrium which are prone to mechanical (not conceptual) errors, or both.

While equilibrium performance increases with age, there is also a substantial innate component: some of our youngest participants play perfectly from the first trial whereas some of our oldest participants do not go beyond one step of reasoning. Even though there is some evidence of learning, repeated exposure is ineffective at bringing participants to play Nash. Finally, performance increases significantly between 8 and 12 years of age and stabilizes afterwards, suggesting that most of what is needed to solve dominance solvable games is acquired by the end of elementary school. Interestingly, most students acquire complex mathematical skills during adolescence. Our observations suggest that this extra knowledge does not translate into better strategic decision making.

The trajectory in elementary school is consistent with theories in developmental psychology that address logical thinking. A major component of the game is the ability to form a hypothesis about the behavior of other players and use this information to select a best response. Existing research shows that by 7 years of age children may think ahead and form correct anticipations (Eliot et al., 1979; Tecwyn et al., 2014). This means that our 3^{rd} graders are already equipped with some of the faculties that are necessary to play the equilibrium. Children have also been shown to develop inductive logic between the ages of 8 and 12 (Feeney and Heit, 2007; Rhodes et al., 2008). This ability is also an essential component for equilibrium behavior, which explains the significant increase in performance in this age range.

A major unexpected feature is the lack of improvement beyond elementary school. Indeed, the ability to perform hypothetical and counterfactual thinking is known to develop throughout middle school (Piaget, 1972; Rafetseder et al., 2013). We therefore anticipated an increase in equilibrium play during middle school and even beyond. While surprising, the result is consistent with the behavior of children in a related (though significantly simpler) two-person beauty contest game (Brocas and Carrillo, 2020a), where we show that equilibrium behavior increases for children between 5 and 10 years of age and stabilizes afterwards. It also complements the works by Harbaugh et al. (2001) and Brocas et al.

(2019) that show a similar trajectory in the development of rational decision-making. Also, the smooth increase during elementary school together with the significant heterogeneity within each age group is consistent with neo-piagetian theories of cognitive development. These reject the concept of strict stages of development and, instead, emphasize individual differences (Morra et al., 2012), sometimes suggesting that many logical abilities are present much earlier than was thought previously (Goswami, 2002). From a developmental research perspective, our study provides additional evidence that stages of development are not fixed or discontinuous. Even though each child might develop abilities in stages and through milestones, heterogeneity in development acts as a smoothing factor of the aggregate developmental trajectory.

Our data also reveals an unexpected gender difference. We do not have an interpretation for this result because sophistication in our game is a logical ability. It is hard to link this finding to the research on gender, cognition and IQ since that literature provides inconsistent results (Lynn and Irwing, 2004; Reynolds et al., 2008). At the same time, gender differences have been observed in beauty contest games as a function of context and incentives (Cubel and Sanchez-Pages, 2017). It is possible that the school environment promotes female confidence, either through specific initiatives (as in LILA) or through general non discriminatory school polices and practices (as in KING). This and the fact that females are often more self-disciplined, which has been shown to explain higher GPA compared to males (Duckworth and Seligman, 2006), may lead them to engage in more steps of reasoning. Further research on this topic would be very enlightening.

Last, subjects with a self-reported preference for science have a significantly higher level of sophistication. Differences across schools and across tracks are also associated with differences in sophistication. Even though it is impossible to link differences across topic inclination, schools and tracks to differences in cognitive ability with a formal test, we view this result as consistent with studies showing a relationship between cognition and performance in games (Brañas-Garza et al., 2012; Gill and Prowse, 2016; Proto et al., Forthcoming; Fe et al., 2020). At the same time, other factors such as differences in class size, bilingualism, or underlying socio-economic variables may have also played a direct or indirect role.

We believe that studying the development of strategic sophistication and its differences across populations is key to understand detrimental decision-making in social settings early in life. Many decisions made by children (and adults) require anticipation of future outcomes and backward induction reasoning. They also often involve second and third parties with their own motivations and incentives. Adolescent are particularly exposed to situations in which strategic sophistication is crucial to avoid wrong decisions. Examples include engaging in risky activities, such as accepting drugs from peers or engaging in unprotected sex. Also, with the development of the Internet, naive users are often preyed upon, asked to provide personal information or tricked into making harmful decisions. Information deliberately intended to deceive young minds also circulates through social media. Making correct decisions in such environments requires understanding the intentions of others and anticipating the consequences of following their advice or opinions. More generally, children and adolescents are gradually discovering the dangers hiding behind social interactions and need to come equipped to detect them, assess them and navigate around them. We conjecture that failures in these abilities are closely related to under-developed logical abilities, and we predict that the level of sophistication of an individual detected through a simple task matches their behavior in social settings.

Last, we have observed less strategic sophistication among students in the challenged track, which also corresponds to children living at or below the federal poverty line. Statistically, those students are experiencing (or will experience) higher rates of detrimental outcomes, such as low academic performance (McFarland et al., 2017), risky behavior (Kirby, 2002) and criminality (Thornberry and Krohn, 2001). The causality between poverty, under-developed cognition and poor decision making is still not well-understood. However, the link suggests that designing an educational environment that promotes the development of logical abilities in general and strategic sophistication in particular is likely to have a positive impact on the outcomes of children and teens living in poverty. For instance, even though enhancing strategic sophistication among those who face economic hardship will not affect their choice sets, it may improve their ability to pick options in those sets.

References

Ayala Arad and Ariel Rubinstein. The 11-20 money request game: A level-k reasoning study. *American Economic Review*, 102(7):3561–73, 2012.

Ellen Bialystok. Consequences of bilingualism for cognitive development. *Handbook of bilingualism: Psycholinguistic approaches*, pages 417–432, 2005.

Pablo Brañas-Garza, Maria Paz Espinosa, and Pedro Rey-Biel. Travelers' types. *Journal of Economic Behavior & Organization*, 78(1-2):25–36, 2011.

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

Isabelle Brocas and Juan D Carrillo. Self-serving, altruistic and spiteful lying in the schoolyard. USC Working Paper, 2019.

Isabelle Brocas and Juan D Carrillo. The evolution of choice and learning in the two-person beauty contest game from kindergarten to adulthood. *Games and Economic Behavior*, 120:132–143, 2020a.

Isabelle Brocas and Juan D Carrillo. Introduction to special issue "understanding cognition and decision making by children." studying decision-making in children: Challenges and opportunities. *Journal of Economic Behavior & Organization*, 179:623–637, 2020b.

Isabelle Brocas, Juan D Carrillo, Stephanie W Wang, and Colin F Camerer. Imperfect choice or imperfect attention? understanding strategic thinking in private information games. *Review of Economic Studies*, 81(3):944–970, 2014.

Isabelle Brocas, Juan D Carrillo, and Niree Kodaverdian. Altruism and strategic giving in children and adolescents. USC Working Paper, 2017.

Isabelle Brocas, Juan D Carrillo, and Ashish Sachdeva. The path to equilibrium in sequential and simultaneous games: A mousetracking study. *Journal of Economic Theory*, 178:246–274, 2018.

Isabelle Brocas, Juan D Carrillo, T Dalton Combs, and Niree Kodaverdian. The development of consistent decision-making across economic domains. *Games and Economic Behavior*, 116:217–240, 2019.

C Monica Capra, Jacob K Goeree, Rosario Gomez, and Charles A Holt. Anomalous behavior in a traveler's dilemma? *American Economic Review*, 89(3):678–690, 1999.

Timothy N Cason and Charles R Plott. Misconceptions and game form recognition: Challenges to theories of revealed preference and framing. *Journal of Political Economy*, 122(6):1235–1270, 2014.

Gordon J Chelune and Ruth A Baer. Developmental norms for the wisconsin card sorting test. *Journal of clinical and experimental neuropsychology*, 8(3):219–228, 1986.

Jingnan Chen, Daniel Houser, Natalia Montinari, and Marco Piovesan. Beware of popular kids bearing gifts: A framed field experiment. *Journal of Economic Behavior & Organization*, 132:104–120, 2016.

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

Suparna Choudhury, Sarah-Jayne Blakemore, and Tony Charman. Social cognitive development during adolescence. *Social cognitive and affective neuroscience*, 1(3):165–174, 2006.

Miguel Costa-Gomes, Vincent P Crawford, and Bruno Broseta. Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 69(5):1193–1235, 2001.

Miguel A Costa-Gomes and Vincent P Crawford. Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review*, 96(5):1737–1768, 2006.

Vincent P Crawford, Miguel A Costa-Gomes, and Nagore Iriberri. Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*, 51(1):5–62, 2013.

María Cubel and Santiago Sanchez-Pages. Gender differences and stereotypes in strategic reasoning. *The Economic Journal*, 127(601):728–756, 2017.

Simon Czermak, Francesco Feri, Daniela Glätzle-Rützler, and Matthias Sutter. How strategic are children and adolescents? experimental evidence from normal-form games. *Journal of Economic Behavior & Organization*, 128:265–285, 2016.

Angela Lee Duckworth and Martin EP Seligman. Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of educational psychology*, 98(1):198, 2006.

John Eliot, Kenneth Lovell, C Mitchell Dayton, and BF McGrady. A further investigation of children's understanding of recursive thinking. *Journal of Experimental Child Psychology*, 28(1):149–157, 1979.

Eduardo Fe, David Gill, and Victoria Prowse. Cognitive skills, strategic sophistication, and life outcomes. Technical report, Working Paper, 2020.

Aidan Feeney and Evan Heit. *Inductive reasoning: Experimental, developmental, and computational approaches*. Cambridge University Press, 2007.

Amanda Friedenberg, Willemien Kets, and Terri Kneeland. Is bounded rationality driven by limited ability? *Unpublished manuscript, Arizona State Univ., Tempe, AZ. http://u. arizona. edu/ afriedenberg/car. pdf*, 2018.

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

Usha Goswami. Inductive and deductive reasoning. *Blackwell Handbook of Childhood Cognitive Development*, pages 282–302, 2002.

William T Harbaugh, Kate Krause, and Timothy R Berry. GARP for Kids : On the Development of Rational Choice Behavior. *The American Economic Review*, 91(5): 1539–1545, 2001.

Eric J Johnson, Colin Camerer, Sankar Sen, and Talia Rymon. Detecting failures of backward induction: Monitoring information search in sequential bargaining. *Journal of Economic Theory*, 104(1):16–47, 2002.

Douglas Kirby. Antecedents of adolescent initiation of sex, contraceptive use, and pregnancy. *American journal of health behavior*, 26(6):473–485, 2002.

Terri Kneeland. Identifying higher-order rationality. *Econometrica*, 83(5):2065–2079, 2015.

LAUSD. Lausd gifted/talented programs. `https://achieve.lausd.net/domain/222`, 2020.

JA List, R Petrie, and A Samek. How experiments with children can inform economics. Working paper, 2018.

Richard Lynn and Paul Irwing. Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, 32(5):481–498, 2004.

Joel McFarland, Bill Hussar, Cristobal de Brey, Tom Snyder, Xiaolei Wang, Sidney Wilkinson-Flicker, Semhar Gebrekristos, Jijun Zhang, Amy Rathbun, Amy Barmer, et al. The condition of education 2017. nces 2017-144. *National Center for Education Statistics*, 2017.

Patricia H Miller, Frank S Kessel, and John H Flavell. Thinking about people thinking about people thinking about...: A study of social cognitive development. *Child Development*, pages 613–623, 1970.

Sergio Morra, Camilla Gobbo, Zopito Marini, and Ronald Sheese. *Cognitive development: neo-Piagetian perspectives*. Psychology Press, 2012.

Josef Perner and Heinz Wimmer. "john thinks that mary thinks that..." attribution of second-order beliefs by 5-to 10-year-old children. *Journal of experimental child psychology*, 39(3):437–471, 1985.

Jean Piaget. Intellectual evolution from adolescence to adulthood. *Human development*, 15(1):1–12, 1972.

Eugenio Proto, Aldo Rustichini, and Andis Sofianos. Intelligence, personality, and gains from cooperation in repeated interactions. *Journal of Political Economy*, Forthcoming.

Eva Rafetseder, Maria Schwitalla, and Josef Perner. Counterfactual reasoning: From childhood to adulthood. *Journal of experimental child psychology*, 114(3):389–404, 2013.

Matthew R Reynolds, Timothy Z Keith, Kristen P Ridley, and Puja G Patel. Sex differences in latent general and broad cognitive abilities for children and youth: Evidence from higher-order mg-macs and mimic models. *Intelligence*, 36(3):236–260, 2008.

Marjorie Rhodes, Daniel Brickman, and Susan A Gelman. Sample diversity and premise typicality in inductive reasoning: Evidence for developmental change. *Cognition*, 108 (2):543–556, 2008.

Itai Sher, Melissa Koenig, and Aldo Rustichini. Children's strategic theory of mind. *Proceedings of the National Academy of Sciences*, 111(37):13307–13312, 2014.

Matthias Sutter, Claudia Zoller, and Daniela Glätzle-Rützler. Economic behavior of children and adolescents–a first survey of experimental economics results. *European Economic Review*, 111:98–121, 2019.

Emma C Tecwyn, Susannah KS Thorpe, and Jackie Chappell. Development of planning in 4-to 10-year-old children: Reducing inhibitory demands does not improve performance. *Journal of experimental child psychology*, 125:85–101, 2014.

Terence P Thornberry and Marvin D Krohn. The development of delinquency. In *Handbook of youth and justice*, pages 289–305. Springer, 2001.

Esther van den Bos, Mark de Rooij, Sindy R Sumter, and P Michiel Westenberg. Continued development of recursive thinking in adolescence: Longitudinal analyses with a revised recursive thinking test. *Cognitive Development*, 37:28–41, 2016.

Henry M Wellman, David Cross, and Julanne Watson. Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, 72(3):655–684, 2001.

Appendix A. Extended analysis

A1. Other regression analysis

We perform robustness checks on the change of behavior with age. In column 1 of Table 9, we run a Probit regression where the binary dependent variable is whether the individual played Nash in all 18 trials. In columns 2, 3 and 4, we report multinomial regressions where the dependent variable is the individual's type. We use D_0 as the omitted category (we do not perform a regression on D_1 as it contains only one individual). In all regressions, we include the same independent variables as in the OLS regression of Table 4.

	Probit	Multinomial regression		
	Nash	D_2	R	O
Age	0.012***	0.016***	-0.061**	-0.002
	(0.003)	(0.006)	(0.024)	(0.007)
Male	-0.350*	-0.645*	0.290	-0.079
	(0.183)	(0.385)	(0.864)	(0.418)
STEM	0.587***	1.579***	0.035	0.568
	(0.188)	(0.447)	(1.002)	(0.495)
Siblings	-0.139	-0.459	0.538	-0.425
	(0.205)	(0.461)	(1.170)	(0.498)
Constant	-2.065***	-1.436	5.657*	0.819
	(0.479)	(1.021)	(3.024)	(1.096)
Log Likelihood	-142.7	—	—	—
AIC	—	502.3	502.3	502.3

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 9: Probit and Multinomial regressions

The Probit regression supports the existing findings that age, gender and a preference for STEM are indicative of Nash behavior. The regressions on types also yield similar conclusions to the standard OLS. Compared to individuals who only recognize a dominant strategy (D_0), older subjects are more likely to play at equilibrium (D_2) and less likely to play randomly (R). A preference for STEM (and to a lesser extent females) increases the likelihood of equilibrium play but has no effect on playing randomly or the dominant strategy. There are no significant differences between D_0 and O, reinforcing the idea that although O types typically play better than random, they are not very sophisticated either.

A2. Preference for school subjects: STEM vs. Arts & Humanities

The difference in performance as a function of school preference noted in Tables 4 and 9 is both surprising and interesting. To investigate this effect in more detail, we present in Figure 10, the same information as in Figure 2, namely the proportion of equilibrium choices by grade and role, separately for subjects with a preference for STEM (86 subjects) and with a preference for Arts & Humanities (148 subjects).

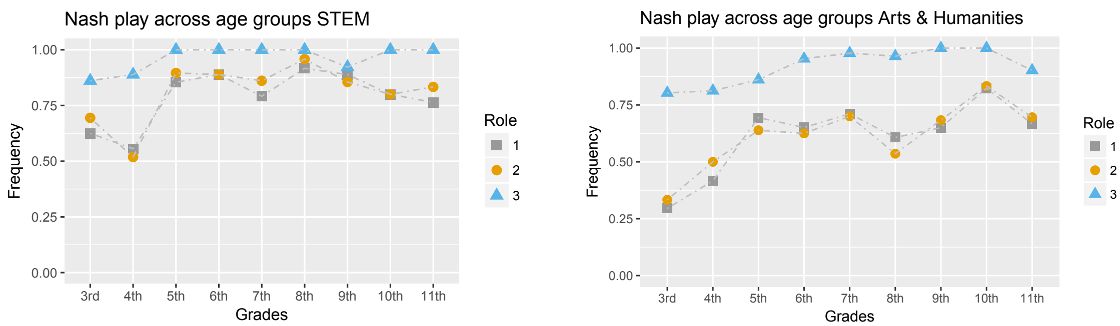


Figure 10: Equilibrium choices by grade, role and favorite school subject

The graph illustrates the difference in performance in roles 1 and 2 across topic preferences. Averaging across grades, subjects who like STEM play 77.9% and 80.4% the equilibrium action in roles 1 and 2 compared to 61.4% and 61.3% for the subjects who prefer Arts & Humanities. These differences are highly significant ($p < 0.001$).

When we run the same OLS regressions as in columns 1, 2 and 3 of Table 4 separately on each subsample, age loses significance for subjects who prefer STEM but not for those who prefer Arts & Humanities. This captures the fact that equilibrium performance of the former is always high independent of age, with the exception of 4^{th} graders (regressions omitted for brevity). Overall, we realize that self-reported preferences partly capture intrinsic taste but they also capture self-perceived ability over topics. Also, while we tried to minimize analytical requirements to understand the game, we may not have fully succeeded. With these caveats in mind, the result nevertheless suggests that a scientific inclination is correlated with equilibrium behavior, and this holds independent of age.

A3. The behavior of non-equilibrium players

Our theory has assumed that an individual who does not unveil the logic of equilibrium play in roles 1 and 2 will choose randomly between the three options. At the same time, and as briefly mentioned in section 2.1, we have introduced non-equilibrium focal objects to minimize the likelihood of spurious equilibrium choices. In this section, we briefly study whether non-equilibrium players follow any discernible strategy.

We define the options as follows. We call *rational* the option chosen by an equilibrium player, *heuristic* the option chosen by a subject who matches the attribute of the extra object, and *alternative* the option chosen by a subject who does neither of the previous two. By definition, *rational* and *heuristic* coincide for role 3. More importantly, the construction of our "focal objects" is such that *rational* and *heuristic* are always different in roles 1 and 2.[25] This means that in roles 1 and 2, there is always one *rational*, one *heuristic*, and one *alternative* option. In Figure 11, we present the proportion of *rational*, *heuristic*, and *alternative* choices in roles 1 and 2 by subjects classified as D_0, R and O.[26]

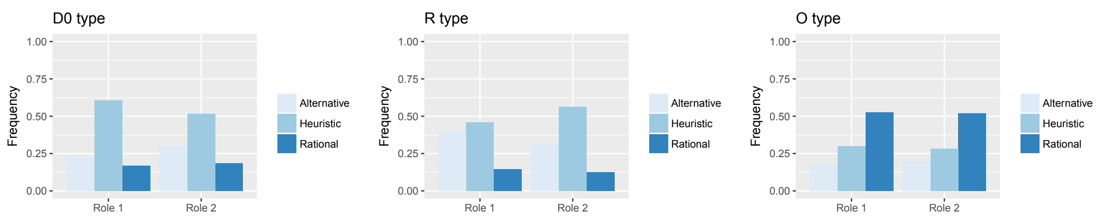


Figure 11: Options chosen in roles 1 and 2

By construction, D_0 and R subjects have chosen the *rational* option at most twice in each role whereas O subjects are likely to have chosen it three or four times (otherwise they would have been classified as one of the other types). It is, therefore, not surprising that *rational* is under-represented in D_0 and R and over-represented in O. Interestingly, all types choose more often *heuristic* than *alternative* in both roles. These differences are significant for D_0 ($p < 0.001$) and for O in both roles ($p < 0.04$). One interpretation of this finding is that subjects who recognize a dominant strategy and only a dominant strategy (D_0) erroneously apply the same logic to other roles as well. Subjects who are less (R) or more (O) sophisticated than D_0 are less prone to this mistake.[27]

A4. Learning

Many of our participants did not grasp the backward induction logic from the outset. However, after playing a few times in different roles, they may have used their behavior in a certain role to deduce what to do in another role. For example, after playing in role 3,

[25] For example, in Figure 1 the *heuristic* option in roles 1 and 2 is the blue square A, which coincides with the extra object and is different from the *equilibrium* options.

[26] We do not include D_2 subjects since, by definition, they have chosen 5 or 6 times out of 6 the *rational* option in both roles.

[27] Notice that in all roles and ages a *rational* choice is more likely than a *heuristic* choice (this is obvious in grades 5^{th} and above but it is also true in 3^{rd} and 4^{th} grade). So, for all roles and ages, even an individual who correctly anticipated the empirical behavior of the age group would find it optimal to play the equilibrium strategy.

they may have understood the dominant strategy in that role and used it to best respond in role 2. If a significant fraction of subjects are in this category, a classification method based on the entire game may be misleading or incomplete.

To address changes in behavior during the experiment, we present in Figure 12 the fraction of equilibrium choices by grade and role (as in Figure 2), separated between the first and last nine trials of the game.

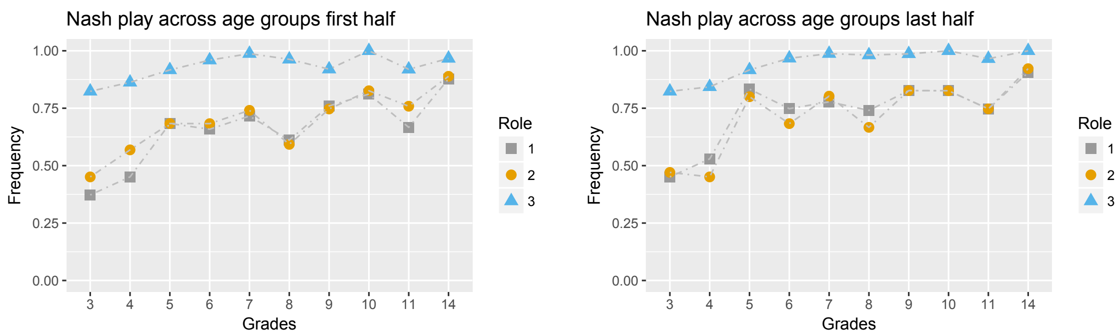


Figure 12: Equilibrium choices in first (left) and last (right) half of trials

We notice a small but sustained increase in equilibrium behavior for roles 1 and 2, with the exception of 4^{th} graders. Averaging across all school-age subjects, individuals play in roles 1, 2 and 3 the equilibrium action 68.5%, 71.6% and 93.6% of the time in the first half of the experiment compared to 75.5%, 74.2% and 95.4% in the second half. The difference is highly significant for role 1 ($p < 0.001$) and marginal for roles 2 ($p = 0.075$) and 3 ($p = 0.052$).

We next perform a similar classification exercise of types as previously, separately in each subsample. With fewer observations, the classification is bound to be more inaccurate. We labeled the behavior in half of the trials of a role "equilibrium" if all 3 observations were consistent with theory and "non-equilibrium" otherwise (0, 1 or 2 out of 3 observations consistent with theory). For each subsample $X \in \{F, L\}$ (where F is the first half and L is the last half), we considered the same types as before: R^X, D_0^X, D_1^X and D_2^X. The remaining subjects are denoted O^X. Table 10 reports the type of subject in the first and last half of the experiment. For this analysis, we focus on LILA population.

The type of two-thirds of our subjects does not change between the first and last half of the trials. Among those who change types, 42% learn to play the equilibrium in the second half (32 subjects) for 8% who play equilibrium in the first half but not the second (6 subjects). Also, 18% improve from R^F to D_0^L (14 subjects) for 8% who decrease from D_0^F to R^L (6 subjects).

Overall, there is some small evidence of change in behavior during the experiment, and

	O^L	R^L	D_0^L	D_1^L	D_2^L
O^F	3	1	2	0	7
R^F	2	12	14	0	0
D_0^F	9	6	48	2	13
D_1^F	0	0	2	3	12
D_2^F	3	0	0	3	92

Table 10: Types of school-age subjects in first (F) and last (L) half of trials.

it predominantly takes the form of learning to play closer to equilibrium.

A5. OLS regression of the YOUNG population

We ran OLS regressions similar to Table 4 with the YOUNG population to better understand the determinants of equilibrium choices in roles 2 and 3. The results are summarized in Table 11.

	OLS	
	Role 2	Role 3
Age	0.011***	0.008***
	(0.003)	(0.003)
Male	0.110*	0.023
	(0.060)	(0.058)
STEM	-0.020	0.070
	(0.068)	(0.065)
Siblings	0.095	-0.010
	(0.067)	(0.064)
Constant	-0.772***	0.017
	(0.243)	(0.233)
Observations	105	105
Adj. R^2	0.150	0.058

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 11: OLS regression of equilibrium choices in roles 2 and 3 for the YOUNG population

Confirming our previous findings, age is a major determinant of equilibrium behavior in both roles. Contrary to our main population, a preference for STEM is not indicative of equilibrium choice, maybe because some of these participants are too young to have clearly established inclinations. Males perform marginally better than females.

Appendix B. Instructions and Quiz

B1. Instructions

Hi, everyone. Today we are going to play a few games. In all the games, you will earn points that will be placed in your virtual wallet.

[For subjects in grades 6 and above] At the end of the experiment you will be paid 1 cent for each point you obtained with an Amazon gift card. You will get several hundred points, so you will be able to get a nice gift card.

[For subjects in grades 3, 4 and 5] At the end of the experiment we will go to the toy shop and you will be able to buy the toys you like with the points you earned.

In all the games, you will play through the tablets. We ask you to not talk and keep your decisions private.

This game is called the "matching game." In this game, you will be playing many times. Each time, you will be playing in groups of 3. The computer will decide with whom you play and you will not know who that is. If you are player 1, you will see a screen like this.

[SLIDE 2]

At the top of the screen, it tells you are player 1. There are 3 large grey pictures on the screen. Yours is the darkest. On this picture, you can read "YOU ARE Player 1". There are 3 objects on this picture. Each object is a colored shape that is marked with a letter. Shapes, colors and letters are all different. You have to select one object by clicking on it and pressing OK. On your screen, you can also see the objects on Player 2's picture and the objects on Player 3's picture. There is also one object outside the 3 pictures.

If you are Player 2, you will see a screen like this.

[SLIDE 3]

This is the same screen as for Player 1 except that your picture is the darkest one in the middle where you can read "YOU ARE Player 2." If you are Player 3, you will see a screen like this.

[SLIDE 4]

Again, this is the same screen as for Players 1 and 2 except that your picture is the darkest one in the middle where you can read "YOU ARE Player 3." All right, now, how do you obtain points?

[SLIDE 5]

In this game Player 1 has to choose an object that has something in common with the object Player 2 chooses. The arrow between the picture of Players 1 and 2 tells you what they need to have in common. In this example, Player 1 needs to choose an object that has the same letter as the object chosen by Player 2. Now what about Player 2?

[SLIDE 6]

Player 2 has to choose an object that has something in common with the object that Player 3 chooses. The arrow between the picture of Players 2 and 3 tells you what they need to have in common. In this example, Player 2 needs to choose an object that has the same color as the object chosen by Player 1. What about Player 3?

[SLIDE 7]

Player 3 has to choose an object that has something in common with the object that is outside the pictures. In that example Player 3 needs to choose an object that has the same shape as the object outside the pictures. Each time you play, you will know what each player needs to do to win because you all see the same screen. However, when you make a choice, you do not know what objects the others have chosen. If you choose the object that matches what you are asked to match, 40 points will be added to your wallet. If you miss, only 20 points will be added. Is it clear for everyone?

Remember, you will play several times. Sometimes you will be player 1, sometimes you will be player 2 and sometimes you will be player 3. Each time you play, you will play with different people. Also, the shapes, colors, letter and characteristics that you need to match will change. We will not tell you how much you earned each time you played. We will only tell you how many points you have earned in total at the end of the game.

Everybody understands? Let's answer some questions, just to make sure everybody understands. Look at the screen here.

[SLIDE 8]

On your computers are some questions that you have to answer correctly before we start the game. If you need help with the questions, raise your hand and we will come to assist you.

Are you ready to start the game? Remember, you will play several times, always with different partners. The roles, the objects and what you need to match will change. We will tell you at the end how many points you obtained.

Figure 13: Slides 1-8 projected in the main screen

B2. Quiz

Below are the questions that each subject would see on their computer screen. All questions had to be answered correctly before the paid part of the experiment could start.

Figure 14: Quiz included before the paid part of the experiment