Adverse Selection and Contingent Reasoning in Preadolescents and Teenagers *

Isabelle Brocas
University of Southern California and CEPR

Juan D. Carrillo
University of Southern California and CEPR

February 2022

Abstract

We study from a developmental viewpoint the ability to perform contingent reasoning and the cognitive abilities that facilitate optimal behavior. Individuals from 11 to 17 years old participate in a simplified version of the two-value, deterministic "acquire-a-company" adverse selection game (Charness and Levin, 2009; Martínez-Marquina et al., 2019). We find that even our youngest subjects understand well the basic principles of contingent reasoning (offer the reservation price of one of the sellers), although they do not necessarily choose the optimal price. Performance improves steadily and significantly over the developmental window but it is not facilitated by repeated exposure or feedback. High cognitive ability–measured by a high performance in a working memory task–is necessary to behave optimally in the simplest settings but it is not sufficient to solve the most complex situations.

Keywords: developmental decision-making, lab-in-the-field experiment, contingent reasoning, winner's curse.

JEL Classification: C93, D82.

* We thank the members of the Los Angeles Behavioral Economics Laboratory (LABEL) –in particular Sobhana Atluri, Sacha Bazzal, Matt Burr, Irfan Khan and Wen Lu– for their insights and research assistance, to Chris Crabbe for his outstanding programming skills, and to David Gill, Muriel Niederle, Matthias Sutter and Emanuel Vespa for helpful comments. We are grateful to the staff of the Lycée International de Los Angeles –in particular Emmanuelle Acker, Mathieu Mondange and Anneli Harvey– for their support in running the experiment in their school. The study was conducted with the University of Southern California IRB approval UP-12-00528. We acknowledge the financial support of the National Science Foundation grant SES-1851915.

1 Introduction

It has been extensively documented in the experimental literature that individuals have a difficult time making profit maximizing decisions in situations involving contingent reasoning. Examples include the winner's curse in common value auctions (Kagel and Levin, 1986), zero-sum betting (Sonsino et al., 2002) and asset trading for informational reasons (Carrillo and Palfrey, 2011). Researchers have proposed behavioral theories to model the imperfect ability of subjects to understand the relationship between the actions and the information of other players (Eyster and Rabin, 2005; Crawford and Iriberri, 2007; Esponda, 2008; Rogers et al., 2009) and studied lookup patterns to differentiate between choice mistakes and limited attention (Brocas et al., 2014).

One of the most elegant formalizations of contingent reasoning is the "acquire-a-company" game (Samuelson and Bazerman, 1985; Ball et al., 1991), where a buyer makes a take-it-or-leave-it offer to a seller for a company. The company's value v is privately known to the seller and is worth a premium to the buyer. This model is analogous to a lemon's problem (Akerlof, 1970) and, under some parametric assumptions, information asymmetry results in no exchange despite the potential welfare gains from trade. The experimental literature has shown that individuals have considerable difficulties in finding the optimal strategy (Grosskopf et al., 2007; Bereby-Meyer and Grosskopf, 2008). Since the setting involves multiple players, beliefs about the behavior of others play a key role in the buyer's choice. Also, computing a conditional expectation requires some relatively challenging bayesian computations that are known to be difficult for many individuals. Overall, this literature has been successful in identifying some of the difficulties faced by participants. However, the complexity of the environment has been a handicap to achieve a full understanding of the reasons for the observed departures.

To further advance this comprehension, the innovative experiments by Charness and Levin (2009) and Martínez-Marquina et al. (2019) (from now on [CL] and [MNV]) strip down the problem to its very essence. [CL] automatize the seller's decision problem and restrict the company's worth to only two possibles values. The paper shows that even in this extremely simplified setting, 40% to 70% of subjects still deviate from the theoretical prediction.[1] [MNV] decompose the problem into what they call "computational complexity" and "loss of power of certainty" to study the importance of each effect in the difficulty to perform contingent reasoning. To achieve this decomposition, they add a deterministic version where the buyer faces two companies, one of each value, but it is required to make the same offer to both. The authors show that equilibrium compliance in

[1] Deviations are still significant when participants play in teams (Casari et al., 2016; Cooper and Sutter, 2018). Also, the response is stronger to adverse selection than to advantageous selection (Ali et al., 2021).

the deterministic version is significantly higher than in the probabilistic one. Equilibrium play increases further if buyers are educated first, by playing a number of trivial rounds against only one company of known value. However, and despite all these improvements, the proportion of individuals who play optimally remains relatively low.

The relatively low compliance with equilibrium play in [MNV] suggests that the probabilistic and deterministic versions of the game pose *qualitatively* similar challenges to the reasoning of players. This is natural if we go back to the very definition of contingent (or conditional) reasoning. Conditional statements take the form "if contingency A is true, then consequence B will follow". In the acquire-a-company game, contingent reasoning is required to determine the kind of offer the seller would accept given their private information (the contingency). To conclude about an optimal offer, the buyer must take the output of contingent reasoning for each possible contingency and use it as a starting point to reason about the best course of action. This step requires recursive thinking. It involves noting that each contingency leads to a different response and resolving these contradictions to pick the best overall. Obviously, a probabilistic setting in which A may or may not be true, because of uncertainty around A, is the prime example of contingent reasoning. However, the same type of reasoning is needed to solve the deterministic problem. Contingent reasoning must be applied to each contingency (which now is always true) and recursive thinking must resolve the same contradictions between accepted offers across sellers. The deterministic formulation only facilitates the representation of contingencies.[2] The question therefore is: why do people find it so difficult to draw conclusions from hypothetical statements (whether true sometimes or always) and act in their best interest?

The goal of the present paper is to provide an investigation of the development of contingent reasoning and the cognitive abilities that facilitate optimal behavior. To this purpose, we study the evolution of behavior in the acquire-a-company game from preadolescence to young adulthood (11 to 17 years old). The above mentioned research shows that contingent reasoning is challenging for educated adults. We want to determine if this skill takes time to acquire, is invariant to age or is lost during adolescence. We also want to find out which features make contingent thinking so complex. In particular, we are interested in studying at which age (if any) are individuals capable of learning from their mistakes. Finally, but importantly, we determine if the ability to perform well in this task is related to known cognitive abilities critical to complex reasoning.

We consider the deterministic version of [MNV] (itself based on [CL]) which reduces

[2]This is consistent with research that reports better comprehension and inferences in all age groups when problems are formulated in natural frequencies as opposed to probabilities (McDowell and Jacobs, 2017).

contingent reasoning to its core logical components, deprived of the power of certainty.[3] We also simplify the task further in three dimensions. First, we present a graphical, story-based version. Limiting the analytical requirements ensures comprehension and helps retaining the interest of our youngest participants without affecting its essence. Second, we consider an additive instead of a multiplicative buyer's premium ($v + x$ with $x > 0$ instead of αv with $\alpha > 1$), which further facilitates numerical calculations. Third, we provide feedback after each round (whether none, one or both companies are acquired and, if they are, the net payoff of each transaction). This allows us to study initial behavior as well as learning. At the end of the experiment, we ask participants to complete a working memory task to assess the contribution of cognition to optimal behavior in the game.

The reader may find it adventurous to study developmental decision-making in settings that are difficult for educated adults. We should notice, however, that existing research using indirect (Harbaugh et al., 2001) as well as direct (Brocas et al., 2019) tests of transitivity has demonstrated that by the end of elementary school, children have achieved a level of GARP consistency and transitivity of preferences comparable to that of adults. Furthermore, our study on dynamic games of complete information (Brocas and Carrillo, 2021b) shows that equilibrium behavior increases significantly with age up until middle school and stabilizes afterwards. These results suggest that, by 10 years of age, children satisfy the basic axioms of rationality and are capable of performing backward induction. In other words, they are equipped with the cognitive tools required to understand decision-making problems and evaluate options logically. Naturally, this is a necessary but not sufficient condition for optimal behavior. The question we address is whether, at this young age, they are able to transform this documented analytical ability into payoff-maximizing choices.

Our main findings are the following. First, the vast majority of our subjects–including the youngest ones–show a solid understanding of the basic principles of (deterministic) contingent reasoning. Indeed, 75% (6th grade) to 97% (10th grade) of participants in a grade offer "reasonable" prices in every round, namely, one of the seller's values. No subject offers the average value of the two sellers. Second, the fraction of individuals who submit the optimal price in every round increases steadily and significantly with age in the entire window of observation, from 11% in 6-7th grade to 50% in 11th grade, and all the way to 70% in the control population. We conjecture that our simple narrative and graphical design may have contributed to an increased comprehension of the problem

[3] In the terminology of [MNV], these elements are combined into a broad "computational complexity" component. Because the reasoning needed to select an offer both in the absence and in the presence of "power of certainty" still requires to think contingently, computational complexity still contains the contingent reasoning component.

relative to the previous literature.[4] Third, comprehension of the simple one-seller problem is predictive of optimal pricing in every age group. It suggests that a main challenge in problems involving contingent thinking is the difficulty to understand basic aspects of the problem, even more than the ability to maximize profits once these aspects are understood. Finally, performance in the working memory task highly correlates with the basic ability to realize which prices are potentially optimal (one of the sellers' values) and, to a lesser extent, with the more subtle ability to optimally discriminate between the two. It suggests that working memory alone accounts for some but not all the variance in reasoning abilities. It also raises a more general question: is working memory a key cognitive ability which is necessary for (and predictive of) contingent reasoning not only in children and adolescents but also in adults? Our tentative answer would be "yes", but future research in adults should explore this correlation.

2 Experimental design

The paper studies a graphical, simplified, deterministic version of the acquire-a-company game ([CL] and [MNV]) in a population of preadolescents and teenagers (11 to 17 years old). This pool of individuals presents some methodological challenges. We follow the guidelines proposed by Brocas and Carrillo (2020a) to address them. In particular, we simplify the procedures given the participants' limited attention, we present the task in a simple and attractive way, and we include a benchmark adult comparison group.

Participants. We recruited 261 participants (133 females) from 6th to 11th grade at the Lycée International de Los Angeles (LILA), a French-English bilingual private school in Los Angeles.[5] For comparison, we ran the same experiment with a control population of 71 USC college undergraduates (U) using the same procedures. Table 1 reports the distribution of participants by grade and approximate age.[6]

Notice that with some exceptions (e.g., Cobo-Reyes et al. (2020)), studies with children usually 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 should be taken with caution. In our case, the majority of students at LILA are from caucasian

[4] Visual representations have been shown to improve accuracy in inference problems (Brase, 2014; Binder et al., 2015).

[5] Students from 12th grade did not participate in the study because they were preparing for national french exams.

[6] We also had the opportunity to conduct the same experiment with two small samples: 8 math teachers at LILA and 11 master students at USC. In Appendix B we report some summary conclusions of these two populations. There are fewer high schoolers in the sample due to some testing constraints but mostly due to a recent expansion of the school that has attracted a larger cohort of middle schoolers.

	LILA						USC
Grade	6th	7th	8th	9th	10th	11th	U
Age	11-12	12-13	13-14	14-15	15-16	16-17	18-23
Participants	55	45	62	36	33	30	71

Table 1: Number of participants by grade

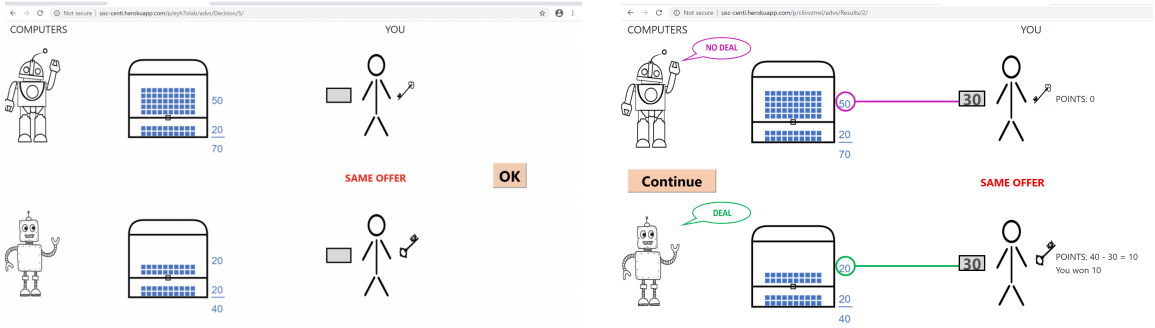
families of upper-middle socioeconomic status. After graduation, most of them attend well-ranked colleges in Europe, Canada and the US (including USC and schools in the UC system). Overall, and despite some differences (nationality, family background, size of peer group, etc.), we believe the two populations are a reasonable match. Also, while the pool has the disadvantage of not being representative of the US population (which raises the issue of external validity), it also has some advantages. In particular, it is homogenous, making it possible to perform meaningful age comparisons. Indeed, as shown in previous research (Charness et al., 2019; Brocas and Carrillo, 2021b), the strategic behavior of children is highly dependent on a variety of economic and demographic characteristics. Pooling participants from different schools could potentially introduce confounds that hide any developmental trajectory. We avoid these confounds by recruiting children from the same school, who follow the same curriculum, and come from similar social and economic backgrounds.[7] We also collect information about gender and number of siblings to capture potential remaining sources of individual heterogeneity. Finally, there is no self-selection within the school, as only 3 students in the entire school opted out of the study. The participation rate was 87% due to school absences and testing constraints on the days of the experiment.

Procedures. We ran 33 and 8 sessions at LILA and USC with 6 to 12 participants each. Sessions at LILA were run in classrooms during school hours with individual partitions to preserve anonymity. Sessions had a mix of male and female participants always from the same grade. Sessions at USC were run at the Los Angeles Behavioral Economics Laboratory (LABEL) in the Department of Economics at USC. Procedures were *identical* in both cases. The experiment was programmed in 'oTree' (Chen et al., 2016a) and implemented on touchscreen SurfacePro PC tablets through a wireless closed network. Due to space and time constraints, we sometimes ran two sessions simultaneously.

The Treasure Game. In developmental studies, it is of paramount importance to provide a simple, graphical interface and a story which is sound, accessible and appealing to a young population. With this goal in mind, we developed the following narrative.

[7] Naturally, it would be even better to have a larger sample size, with children from multiple schools and diverse characteristics.

Each of two computer robots own a treasure chest. Chests have an upper and a lower compartment. Robots do not have the key for the lower compartment so the value of the chest to them is the amount of points in the upper compartment. You, the participant, own the keys of the lower compartments. This means that the value of the chests to you is the amount of points in both the upper and lower compartments. You can purchase one or both chests. Robots are preprogrammed to accept any offer equal to or greater than the value in the upper compartment of the chest. Your task is to choose which offer to make given one condition: it has to be the same for both robots.[8] Figures 1a and 1b provide screenshots of the game.[9]



(a) Initial choice **(b)** Feedback

Figure 1: Screenshots of the Treasure Game

Basic theory. Let v_i ($\in \{\underline{v}, \overline{v}\}$) be the value of the chest for one of the sellers (points in the upper compartment), with $0 < \underline{v} < \overline{v}$. Let x (> 0) be the extra value for the buyer (points in the lower compartments of the chests which, for simplicity, are constrained to be the same in both chests). Under common knowledge of $(\underline{v}, \overline{v}, x)$, if the buyer offers a price p his payoff $\pi(p)$ is:

$$\pi(p) = \begin{cases} 0 & \text{if} \quad p < \underline{v} \\ (\underline{v} + x - p) & \text{if} \quad \underline{v} \leq p < \overline{v} \\ (\underline{v} + x - p) \ + \ (\overline{v} + x - p) & \text{if} \quad \overline{v} \leq p \end{cases}$$

[8] Formally, if the participant inputs an offer for one robot, the offer for the other robot gets automatically populated with the same number.

[9] During the design phase, we read the instructions to a small set of students (who obviously did not participate in the experiment) and noticed that some children had a hard time understanding asymmetric valuations. A chest with different compartments and a privately owned key was a natural way to explain it. Using robots instead of humans emphasized the deterministic nature of the decision rule of opponents.

This means that the optimal price p^* and the corresponding payoff $\pi(p^*)$ are:

$$p^* = \begin{cases} \underline{v} & \text{if} \quad \underline{v} + x \leq \overline{v} \\ \overline{v} & \text{if} \quad \underline{v} + x \geq \overline{v} \end{cases} \quad \text{and} \quad \pi(p^*) = \begin{cases} x & \text{if} \quad \underline{v} + x \leq \overline{v} \\ x + \underline{v} + x - \overline{v} & \text{if} \quad \underline{v} + x \geq \overline{v} \end{cases}$$

Notice that, while the logic behind the optimal price is identical in our setting than under the standard multiplicative structure, the analytical formulation is simpler. This simplification allows us to screen out computational mistakes and concentrate on cognitive mistakes that are not based on the ability to manipulate quantities. Also, the pre-announced strategy of the robots transforms the game into an individual decision making problem (as pioneered by [CL]), but it still retains many features of a strategic game.[10]

Timing. The experiment involved the following six steps. First, we read the instructions. Second, to ensure understanding of the rules of the game, we implemented a computerized, four-question, multiple choice quiz that every participant had to answer correctly before moving on.[11] Third, we moved to the paid part of the experiment. To further facilitate comprehension, we followed [MNV] and started with three (paid) "warm-up" rounds of the Treasure Game, where subjects played against only one robot, and only one chest with values v and x in the upper and lower compartments, respectively.[12] This was followed by twelve rounds of the core game with two robots, see Figure 1a.[13] The only difference across rounds were the values in the upper and lower compartments of the chests $(\underline{v}, \overline{v}, x)$. Values were chosen in a way that the optimal price p^* switched every two rounds between $\underline{v}$ and $\overline{v}$ (see Table 7 in Appendix A4). After each round, participants received feedback on whether the offers were accepted by the robots (no deal or deal) and the net points accumulated, zero under no deal and value of the chest minus price offered $(v_i + x - p)$ under deal, as reflected in Figure 1b. This section was self-paced, with no limit on the time participants could take. It was programmed in a way that subjects had to wait for others to finish only at the end of the 15 rounds. Fourth, there was a two-part, non-incentivized questionnaire. The first part of the questionnaire was one round identical to the main treasure game, except that we also elicited the participants' confidence

[10] As emphasized in the surveys by Sutter et al. (2019) and List et al. (2018), most economic experiments with children and adolescents focus on rationality of choices, time preferences, risk preferences and social preferences. There are a few exceptions of games of strategy studies (e.g., Murnighan and Saxon (1998); Harbaugh and Krause (2000); Sher et al. (2014); Czermak et al. (2016); Chen et al. (2016b); Fe et al. (2020); Brocas and Carrillo (2020b, 2021b, forthcoming)).

[11] If a participant missed one or more answers, a warning sign would appear stating "not all answers are correct, please try again". We provided individual coaching for individuals who struggled with some question (around 10%-15% of our participants).

[12] Our initial idea for the warm-up rounds was to have two robots but allow two different prices. Given the success of the one-robot treatment in [MNV], we decided to follow their design.

[13] Notice that the screen reports not only the values in each compartment but also the total. Again, the goal is to minimize departures from optimality that are due exclusively to computational mistakes.

in their response.[14] The second part of the questionnaire was one round analogous to the treasure game but framed in the standard probabilistic setting, also with a question regarding the participants' confidence. The goal of this section of the experiment was to obtain a quick, simple (but obviously crude) measure of [MNV]'s power of certainty in our setting.[15] Fifth, we ran a computerized version of the Backwards Digit Span Task (Wechsler, 1949). Participants observed a sequence of digits in their screen (from three to eight) and had to report them in reverse order. Participants would obtain points in a trial only if they reported the entire sequence correctly (see section 4.2 for details). The Digit Span Task is a simple behavioral measure of working memory capacity. Working memory refers to the ability to maintain and manipulate information during a cognitive activity (Baddeley and Hitch, 1974) and it is critical to reasoning tasks, in particular tasks that require several reasoning steps and the manipulation of multiple pieces of information (Engle et al., 1999). Since the treasure game requires working memory and cognitive processing, we conjectured a positive correlation between performance in the Digit Span Task and optimal choice in the game. This section was also self-paced. Sixth and last, we conducted a questionnaire regarding, age, gender, siblings and favorite topic at school. In Appendix A, we present a transcript of instructions, quiz, and two-part questionnaire, as well as a full description of the payoffs used in the 15 rounds of the treasure game.

Payoffs. Participants accumulated points during the experiment. Points were converted into money paid immediately at the end of the experiment in cash (USC) or with an amazon e-giftcard (LILA, where cash transfers on premises are not allowed). The conversion rate was \$0.02 per point. There was a \$5 show-up fee paid only to USC students to account for differences in marginal value of money and opportunity cost of time. Average earnings were \$12.5 at LILA and \$14.2 + \$5 show-up fee at USC. The experiment never exceeded 50 minutes (one school period) including instructions and payments.

3 Results

3.1 Aggregate behavior

We first study aggregate choices in the LILA population and compare it with the control USC group. Figure 2 reports the average fraction of optimal play in each grade. We group the rounds in three cases. First, the three warm-up rounds with *one* computer

[14] We asked: "Do you think your offer got you the most possible points?" (Yes / Maybe / No).

[15] To be clear, our study *is not* centered around the comparison between the deterministic and probabilistic versions of the treasure game, as in [MNV]. Based on previous research on adults, we believed that the deterministic version would be already demanding enough for our population. Our core question is the evolution with age of contingent thinking with no uncertainty and the correlates with cognition.

(hereafter referred to as "o-rounds") where the optimal offer is v. Second, the six rounds with two computers where the optimal offer is the *low* value $\underline{v}$ (hereafter referred to as "l-rounds"). Third, the six rounds with two computers where the optimal offer is the *high* value $\overline{v}$ (hereafter referred to as "h-rounds"). However, remember that l- and h-rounds are intertwined in the experiment. For the rest of the paper, we also count as an optimal choice if the participant offers one token above the optimal number ($v+1$ in o, $\underline{v}+1$ in l and $\overline{v}+1$ in h). Indeed, subjects were concerned about the tie-breaking rule. Despite our explanations during the instruction period, it is evident from the answers that some individuals preferred to "play it safe." With a slight abuse of notation, we call $v^+ \in \{v, v+1\}$, $\underline{v}^+ \in \{\underline{v}, \underline{v}+1\}$ and $\overline{v}^+ \in \{\overline{v}, \overline{v}+1\}$.

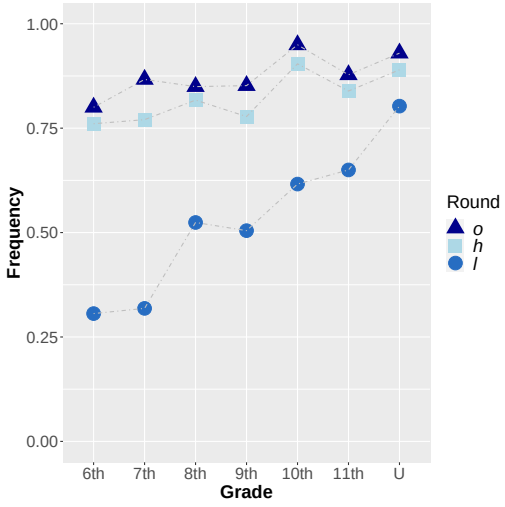


Figure 2: Optimal play by grade

Figure 2 already delivers some interesting conclusions. Optimal choice is very frequent in all school-age grades in o-rounds and h-rounds, between 76.1% and 94.9% of the time, with no significant differences between o and h in any grade except 7th ($p = 0.019$). There are significant differences across grades between 6th-7th grade and 10th grade in both o and h, but they do not survive multiple comparison corrections.

Optimal choice is significantly lower in l-rounds than in h-rounds in all school-age grades ($p < 0.01$ for 6th to 10th grade and $p = 0.020$ for 11th grade), and marginally lower in the control population ($p = 0.048$). In l-rounds, we also observe that participants in 10th and 11th grade perform significant better than participants in 6th and 7th grade ($p < 0.005$ after multiple comparison corrections using false discovery rate controlling procedures). The difference in equilibrium choice between h- and l-rounds is natural, since

the former are intuitively easier than the latter. Indeed, in h-rounds, optimal behavior prescribes offering the minimum price acceptable by both robots ($\overline{v}^+$), which is a very simple strategy. In l-rounds, such instinct must be overcome as a non-profit maximizing strategy, and this extra step of reasoning can be cognitively challenging. Its difficulty may be exacerbated by the fact that choosing $\overline{v}^+$ still provides a positive payoff, so that its sub-optimality might not be put to question.[16] Overall, the results indicate a relatively higher fraction of optimal behavior compared to the previous literature. Also, optimal choice is less frequent when it prescribes foregoing the high value item.

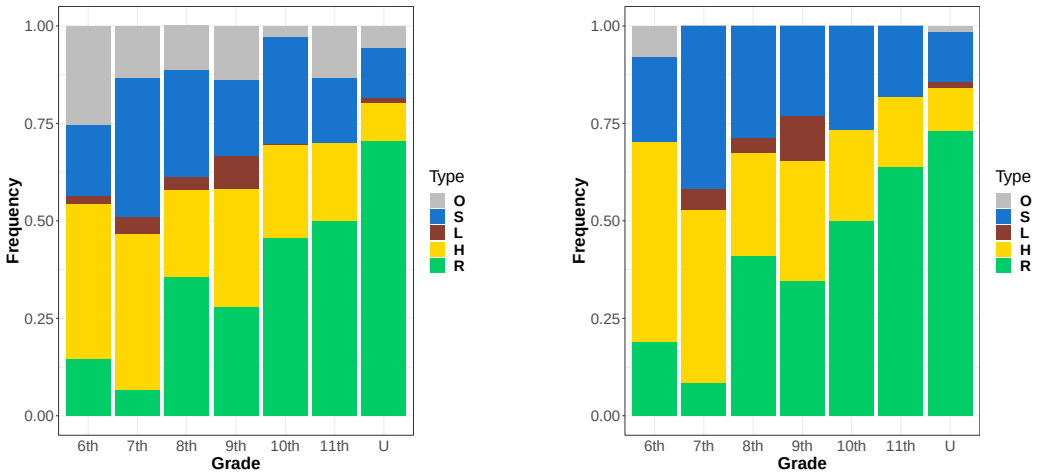
3.2 Individual analysis

Although aggregate behavior is instructive, we are more interested in understanding choice at the individual level. To this purpose, we classify individuals as a function of their behavior in l and h and consider the following types. *Rational* (**R**) is an individual who always offers the optimal price ($\underline{v}^+$ in l and $\overline{v}^+$ in h). *High* (**H**) is an individual who would play optimally if all rounds were h (always offers $\overline{v}^+$). *Low* (**L**) is an individual who would play optimally if all rounds were l (always offers $\underline{v}^+$). *Semi-rational* (**S**) is an individual who understands that the only prices that can be optimal are $\underline{v}^+$ and $\overline{v}^+$ but is not of one of the types above (offers $\underline{v}^+$ and $\overline{v}^+$ in some other proportions). *Other* (**O**) is an individual who offers prices other than $\underline{v}^+$ or $\overline{v}^+$, and has therefore misunderstood the basic principles of the game. To accommodate some minor deviations, we allow for one deviation both in l-rounds and in h-rounds. This means that at least five out of six choices must fall in the corresponding type.[17] Figure 3a reports the proportion of individuals of each type by grade in the entire population (332 participants). Figure 3b reports the same information in the subpopulation of individuals who played correctly all three o-rounds (76.6% of LILA students and 88.7% of USC students for a total of 263 participants).

Even our youngest school-age participants understand remarkably well the basic principles of the treasure game. Indeed, only 3.0% (10th grade) to 25.5% (6th grade) of the subjects offer prices other than $\underline{v}^+$ or $\overline{v}^+$. Furthermore, and as highlighted in Figure 3b, of the 41 type-**O** subjects, the vast majority (37) are individuals who did not play correctly the o-rounds. It means that individuals who understand the basic one-robot problem invariably avoid any price other than v_i^+ in the two-robot problem. The proportion of rational subjects is higher in 8th, 9th, 10th and 11th grade than in 6th and 7th grade (p

[16] To disentangle between different motives for choosing $\overline{v}$ when it is not optimal, one could design variants where (i) participants face three chests and/or (ii) choosing $\overline{v}$ in an l-round results in a negative payoff (formally, by setting $\overline{v} - \underline{v} > 2x$).

[17] Small changes in the deviations allowed have no significant effects on the results. Those robustness checks are omitted for brevity but available upon request.



(a) All subjects **(b)** Subjects with optimal choice in o

Figure 3: Types by grade

< 0.01 after multiple comparison corrections). Consistent with the findings in section 3.1, a significant fraction of individuals (25.9%) choose $\overline{v}^+$ in every round (**H**) whereas very few of them (2.7%) choose the more conservative $\underline{v}^+$ in every round (**L**). Overall, there is a gradual shift from non-optimal v_i^+ prices (**H** and **S**) to optimal v_i^+ prices (**R**) with age. There are also no statistically significant differences between 11th graders and U.

Among the 73 individuals classified as **S**, we observe all types of deviations: 46.6% play optimally in h but not in l, 26.0% play optimally in l but not in h, and the rest do not play optimally in either h or l. We did not find discernible patterns of play among the 41 participants classified as **O**: 1 subject offers prices between $\underline{v}^+$ and $\overline{v}^+$ in 8 rounds, 2 subjects offer prices above $\overline{v}^+$ in all rounds and the remaining 38 subjects offer prices in the whole spectrum. Overall, these individuals appear lost, with no consistent behavior or sign of improvement after feedback.

To our surprise, we found no evidence of dynamic learning by participants within the 12 rounds of the experiment. Learning is a possibility only for **S** and **O** since, by definition, the other types play the same strategy during the entire experiment. If we focus on the last 8 rounds of the experiment, only 4 out of the 73 type **S** subjects are reclassified as **R** and none of the **O** subjects is reclassified as any other type. The absence of learning is consistent with the original findings in [CL] but contrasts with recent results (Ali et al., 2021). It also departs sharply from our recent game theoretical studies with children: our participants learn to play the dominant strategy the second time they play the two-person beauty contest (Brocas and Carrillo, 2020b) and they also coordinate

better in the fair and efficient dynamic outcome the second time they play the repeated stag hunt or the repeated battle of the sexes (Brocas and Carrillo, 2021a). We find this lack of learning especially intriguing, given our deterministic environment. Indeed, since we provide feedback against both sellers, no counterfactual thinking is necessary, and offering $\overline{v}^+$ in an l-round translates into an observed negative payoff against one of the robots. We can think of three (a posteriori) explanations for this result. First, given the numerical values adopted in our game, participants who offer $\overline{v}^+$ in an l-round obtain positive payoffs. This may be enough to avoid triggering suspicion about the sub-optimality of the strategy. Second, absence of learning is easier to understand when a participant offers $\underline{v}^+$ in an h-round, since the choice results in an unrealized gain. Third, a more basic explanation could be that learning is fully concentrated in o-rounds. When the individual reaches the two-robot part of the experiment, they have already made up their mind about the strategy for the remaining of the game, and do not consider revising it. In any case, understanding the circumstances that facilitate learning to perform contingent reasoning is an important area for future research.

Finally, it is informative to contrast our results with [MNV]. The exercise should be taken with caution, not only because of the widely different populations (school-age and college students vs. MTurk workers), but also due to the significant differences in design and procedures.[18] Since their treatment $onevalue_{det}$ is closest to our design (and the one where subjects performed best), we use it as the benchmark for comparison. In that treatment, optimal behavior occurs 47.2% of the time, which is similar to our 10th and 11th graders (47.6%) and significantly lower than our undergraduates (70.4%, p = 0.001). Prices other than v_i^+ occur 26.4% of the time, which is comparable to our 6th graders. It is higher than all other school-age participants grouped together (11.2%, p $<$ 0.001) and also higher than our undergraduates (5.6%, p $<$ 0.001). Finally, the simple one firm setting is solved correctly by 62.6% of subjects, which is significantly smaller than our school-age subjects (76.6%, p $<$ 0.001) as well as our college students (88.7%, p $<$ 0.001). In general, and with the above mentioned important caveat in mind, departures from theory are typically smaller both in our control group and in our older school-age participants than in the existing literature on adults. The comparison highlights the importance of running a control adult group with the same procedures to provide the best possible benchmark. It also validates our choice of focusing on the deterministic problem: while [MNV] convincingly show that probabilistic thinking is a major hurdle for optimal behavior, contingent reasoning is already difficult (and therefore worth of a developmental

[18] Our in-person, graphical instructions and quiz together with our additive formulation and feedback are likely to facilitate understanding. On the other hand, the smaller number of periods and the intertwining of l- and h-rounds could presumably make our game more challenging.

study) in the deterministic version.

3.3 Regression analysis

We perform OLS regressions of choices–where the observation is one individual–to better understand the determinants of optimal play. Since our main interest is the developmental aspect, we consider only the 261 subjects in the school-age population (LILA).[19] In column (1), we report for each individual the proportion of optimal choices in h-rounds ($\overline{v}^+$) and the proportion of optimal choices in l-rounds ($\underline{v}^+$) as a function of the participant's *Age* (in months) at the time of the experiment. We include a dummy variable for l-rounds (to distinguish between performance in those qualitatively different rounds) as well as an interaction term between age and performance in l-rounds. We also add as a control variable the proportion of correct answers in the simple o-rounds, *correct-o*. In column (2), we include dummy variables for gender ($Male = 1$), whether the participant has one or more siblings ($Siblings = 1$), and favorite topic at school to account for analytical inclination ($STEM = 1$), as well as interactions terms of these variables with performance in l-rounds.[20] Finally, column (3) performs the same regression as (2) but only in the subsample of 200 individuals who answered correctly the three initial o-rounds. Presumably, those participants had understood best the core principles of the game (naturally, we do not include the variable *correct-o*). Results are presented in Table 2.

Consistent with the trends in Figure 2, performance in l-rounds is much lower than in h-rounds, and behavior improves with age but only in l-rounds, whether we include control variables (column 2) or not (column 1). The improvement is constant, and males perform better than females but only in the more difficult l-rounds.[21] This is in sharp contrast with our previous research in dominance-solvable games (Brocas and Carrillo, 2021b), where females outperformed males and behavior reached a plateau by middle school. Interestingly, a better performance by males is also reported in [MNV] on a very different population. We found no effect of siblings or preferred school topic. Performing correctly the simplest o-rounds is a very strong predictor of optimal behavior in h- and l-rounds. This is not overly surprising since only individuals who price accurately against one robot are likely to choose v_i^+ against both.[22] Results are very similar when we restrict

[19] We cannot add the USC population since we did not collect their age. Even if we knew their age, we would not want to include them in the regression as it could severely impact the age-trend. We view these subjects as a good control group but not as an extension of the school-age population.

[20] STEM refers to a self-reported preference for either Mathematics or Science/Technology. Consistent with the curriculum of the school, the three other categories offered are Languages, History/Geography and Arts/Music, which we globally refer to as 'Arts & Humanities'.

[21] There is no gender difference in the simplest o-rounds either (regression omitted for brevity).

[22] We performed a robustness check by running additional regressions where *correct-o* is replaced by a

	(1)	(2)	(3)
Age	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)
correct-o	0.169***	0.169***	—
	(0.014)	(0.014)	
l-round	-1.000***	-0.981***	-1.247***
	(0.243)	(0.245)	(0.288)
Age × *l-round*	0.004**	0.003*	0.005**
	(0.001)	(0.001)	(0.002)
STEM	—	-0.008	-0.024
		(0.032)	(0.033)
STEM × *l-round*	—	0.059	0.085
		(0.064)	(0.074)
Male	—	0.020	-0.013
		(0.031)	(0.032)
Male × *l-round*	—	0.173**	0.150*
		(0.062)	(0.074)
Siblings	—	-0.003	-0.001
		(0.037)	(0.037)
Siblings × *l-round*	—	-0.019	-0.023
		(0.081)	(0.094)
Constant	0.193	0.191	0.729***
	(0.121)	(0.125)	(0.113)
Adj. R^2	0.359	0.371	0.289
# observations	522	522	400
# clusters	261	261	200

(standard errors in parenthesis)
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 2: OLS regressions of proportion of optimal choices in *l*-rounds and in *h*-rounds by school-age participants in the entire sample ((1) and (2)) and by the subset of participants who answered correctly the *o*-rounds (3).

attention to individuals with perfect understanding of *o*-rounds (column 3).

In order to better understand the determinants of rational play, we conduct a multinomial logistic regression of the participants' type on their age as well as the previous dummies for gender, siblings, and STEM inclination. The default category in that regression is type **S**. The results reported in Table 3 indicate that the likelihood of being categorized as **R** instead of **S** is positively related to age. It is also more likely when the participant is male and has a preference for STEM.[23]

dummy variable with value 1 for subjects who play all three *o*-rounds correctly. The results are unchanged.

[23]We also created a dummy variable for each possible type (**R**, **H**, **L**, **S**, **O**) and performed independent Probit regressions to explain the probability of being categorized as each of these types on the same

	R	**H**	**L**	**O**
Age	0.028***	-0.003	-0.004	-0.010
	(0.009)	(0.009)	(0.019)	(0.011)
STEM	0.647*	0.212	1.301*	-0.105
	(0.376)	(0.368)	(0.790)	(0.465)
Male	0.813**	-0.352	1.246	0.127
	(0.367)	(0.348)	(0.864)	(0.420)
Siblings	-0.645	-0.579	-0.825	-0.622
	(0.483)	(0.452)	(0.911)	(0.530)
Constant	-4.814***	1.318	-2.180	1.494
	(1.597)	(1.489)	(3.237)	(1.834)
AIC	747.4	747.4	747.4	747.4
# obs.	261	261	261	261

(standard errors in parenthesis)
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 3: Multinomial logistic regression of type on demographic variables (default = type **S**)

3.4 Discussion

Optimal behavior in the treasure game requires several steps. First, to apply conditional reasoning in order to determine the set of offers each seller is willing to accept (above $\underline{v}$ for one seller and above $\overline{v}$ for the other). Second, to understand that the best course of action vis-a-vis each seller is to offer the lowest price in the set, that is, to realize that $\underline{v}$ and $\overline{v}$ are the only candidates for optimal prices. Last, to determine in each round which of these two prices yields the highest payoff, taking into account the output of the conditional reasoning stage. This recursive reasoning, which uses as input the output of the two previous stages, itself requires counterfactual logic ("what would have been my payoff had I instead chosen the other price"). Our results show that the overwhelming majority of participants at all ages (with the exception of some 6th graders) successfully solve the first two steps. By contrast, solving the third step is more challenging, and there is a sustained increase with age in the proportion of individuals who are successful at it.

One reason for the inability to optimally discriminate between the two prices can be computational: participants know what they need to compute but they make algebraic mistakes. Our additive formulation was intended to minimize this possibility by reducing the operations to addition, subtraction and comparison. We also did not constrain the amount of time spent on each round. While computational difficulty may account for a

independent variables. These regressions confirm that the the probability of being categorized as **R** instead of not **R** is correlated with age.

small fraction of mistakes, it is unlikely to be a major driving force for individuals who err frequently, especially those who always choose the same price (types **L** and **H**).

Alternatively, it may just be cognitively difficult to use conditional statements in a recursion that requires counterfactual logic. The improvement of performance with age may correspond to the known development of abstract thinking (counterfactual or conditional) during that period (De Neys and Everaerts, 2008; Rafetseder et al., 2013). As we grow, we become more able to combine logical pieces of reasoning into a broader logical puzzle. The stability of types between the first and second half of the experiment suggests that participants crafted strategies and sticked to them. These strategies contained elements of logic, a strong indication that participants tried to apply a logical argument.

However, the stability of types also shows that participants did not learn or adapted (even mechanically) to the empirical observation of past outcomes. This might have been partly caused by our payoff structure. Indeed, while there is an opportunity cost of choosing suboptimally, payoffs under $\underline{v}$ and $\overline{v}$ are always positive. This makes it less likely to trigger suspicion about the existence of a better alternative.

4 Other analysis

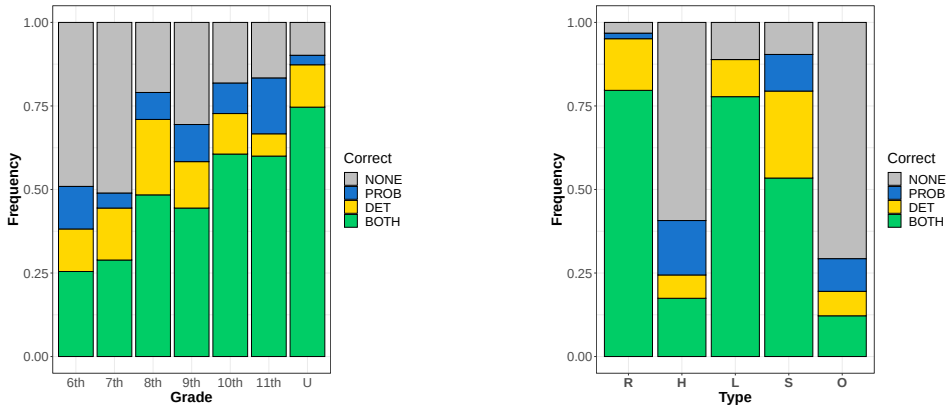
4.1 Deterministic vs. stochastic presentation

Remember that, immediately after the paid rounds, we elicited (non-incentivized) behavior in a deterministic and a probabilistic version of the treasure game.[24] This simple (but obviously imperfect) procedure allows us to obtain a rough measure of the power of certainty. We selected values in such a way that the optimal price is $\underline{v}$ in both cases.[25] Figure 4a reports the fraction of participants in each grade who answered correctly both questions (BOTH), only the deterministic (DET), only the probabilistic (PROB) and none of them (NONE). Figure 4b presents those same fractions as a function of the individual's type in the main experiment.

From Figure 4a, we notice that the evolution with age of individuals who respond correctly to both questions follows the same pattern as the evolution of **R** types in Figure 3a: a sustained increase (in this case, from 25% to 75%) with significant differences between 6th-7th and 10th-11th ($p < 0.030$). Less than one in four students answer correctly exactly one question and, in support of [MNV], the correct answer is statistically more frequent

[24] As explained in Appendix A3, we implement the probabilistic version with the following instructions. "You are matched either with the top computer or the bottom computer, but you do not know which one of them. There is an equal chance it is either of them. However, you have to make your offer before knowing with which computer you are matched."

[25] We assumed (correctly given the results in section 3) that it was more difficult to play correctly in l-rounds than in h-rounds.



(a) Choice by grade **(b)** Choice by type

Figure 4: Choice in the deterministic and probabilistic questions

with the deterministic presentation than with the probabilistic one (14.5% vs. 8.4%, p = 0.020). Figure 4b highlights the tight relationship between choices in the incentivized task and questionnaire: BOTH is highest among **R** followed by **S**. Individuals who always choose $\overline{v}^+$ (**H**) or prices other than v_i^+ (**O**) rarely answer the questions correctly.[26]

Among subjects with an incorrect response to the deterministic question (PROB and NONE), 80.0% offer $\overline{v}^+$ in that question and the rest offer prices in the whole spectrum. Among subjects with an incorrect response to the probabilistic question (DET and NONE), 72.9% offer $\overline{v}^+$ in that question and 9.3% offer $(\underline{v}^+ + \overline{v}^+)/2$. Hence, there is a small but positive "cursedness" effect (offer the average value), which arises only under the probabilistic presentation.

In the questionnaire, we also asked our participants to report if they believed they had made the best possible offer (Yes / Maybe / No), in an attempt to elicit the confidence in their answer. Figure 5 reports their confidence as a function of their choice.

Participants are significantly more confident in the correctness of their response in the deterministic case (left) than in the probabilistic one (right). This is true not only for individuals who answer both questions correctly (79.3% vs. 48.2%, $p < 0.001$), but also for the others (45.2% vs. 23.8%, $p < 0.001$). Again, it supports the argument in [MNV] that uncertainty adds difficulty to contingent reasoning. In our case, probabilistic vs. deterministic implies a small difference in choice but a larger one in confidence.

[26] Type-**L** individuals also answer the questionnaire correctly. However, there are only 9 subjects. Also, it is possible that they use a heuristic (always offer $\underline{v}^+$) that "coincides" with the optimal choice in our two questions.

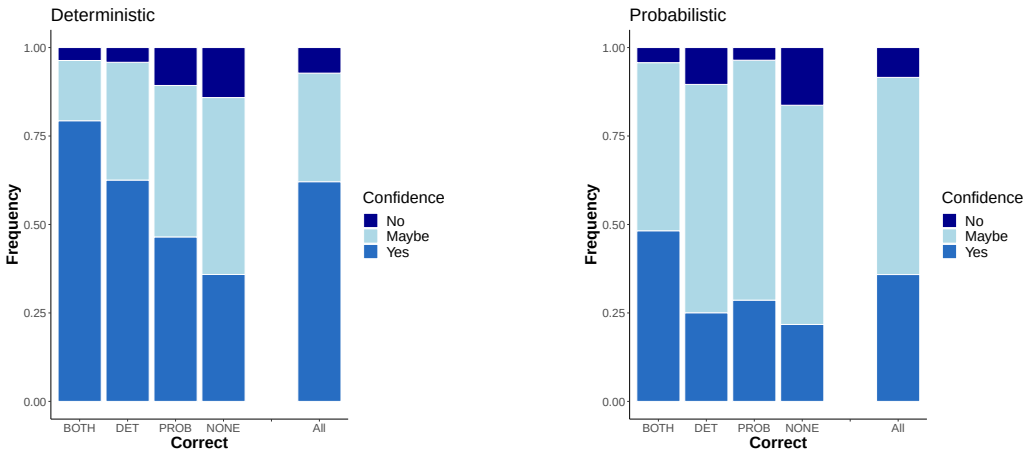


Figure 5: Confidence in correctness of their choice

4.2 Backwards Digit Span Task

The Digit Span Task has been frequently used in Psychology and Neuroscience to measure working memory, the ability to retain and manipulate information for a brief period of time (less than 30 seconds) during a cognitive task.[27] Working memory is recognized as a *specific contributor to fluid intelligence* and the Digit Span Task is administered as a section of intelligence tests in children, as well as in adults, such as the Wechsler's Intelligence Scale (Wechsler, 1949). The task has been rarely used by experimental economists, who favor instead other tasks that provide more general assessments of fluid intelligence, most notably different versions of the Raven's IQ test (see e.g., Brañas-Garza et al. (2012); Gill and Prowse (2016); Proto et al. (2019, 2020); Fe et al. (2020)). Performances in Raven and Digit Span are often correlated. This is not surprising given that solving the logical puzzles in the Raven's test requires working memory among other abilities. We view both tests as complementary and advocate expanding the use of different measures as a rich way to assess cognitive ability. Perhaps one advantage of the Digit Span Task is its narrower focus on one particular primitive ability, which allows us to determine whether working memory is or is not a contributor in the performance of the decision making task. Importantly, the development of abstract, counterfactual and conditional thinking (which our paradigm relies on) has been shown to depend on the ability to maintain and manipulate information in working memory (Handley et al., 2004; Dumontheil, 2014).

[27] For the reader unfamiliar with the Psychology and Neuroscience literatures, it is worth clarifying that working memory does not relate to the traditional definition of memory, that is, the ability to recall events from the past. Instead, it is a cognitive processing ability that involves very different areas of the brain.

Although the original version of this task was verbally administered, recent versions typically use a computer interface. The task exists in two variants: forward-span and backward-span. The backward-span version requires the participant not only to hold the digits longer in their working memory but also to manipulate them because they need to be reordered. It is considered a better measure of working memory capacity (Oberauer et al., 2000). It has also been shown that the ability to perform well on this additional difficulty of the task can be linked to general intelligence (Jensen and Figueroa, 1975).

Figure 6 describes our implementation. Participants observe a sequence of digits in the top of their screen. Digits appear sequentially for 0.5 seconds with an interval of 0.75 seconds between digits. After all digits have appeared, participants must report them in reverse order by consecutively typing one digit in each box from left to right and pressing OK.[28] We start the task with three digits and increase the sequence by one digit every two trials until we reach eight digits, for a total of twelve trials. Participants obtain points in a trial only if they report the entire sequence correctly.

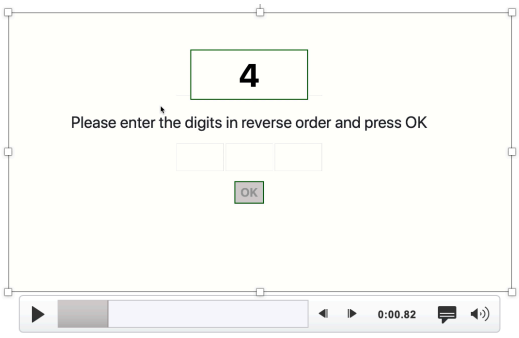


Figure 6: Screenshot of the backwards Digit Span Task

Table 4 reports the average performance in each grade by counting the percentage of trials that the participant performs correctly.[29]

The percentage of correct trials in the school-age population is significantly lower than in the control undergraduate group. While performance in the Digit Span Task is valuable

[28] Participants are obviously not allowed to use any external support (such as writing down the digits). They cannot input the digits as they appear, input them in forward order from right to left, or change a digit after it has been typed. By waiting to press "OK", participants can pace the task and enjoy a break between trials.

[29] It is frequent in Psychology to assess performance by stopping the first time a report is incorrect. Our method is more permissive with mistakes. It also adapts better to our task since all trials are incentivized and performance feedback is obtained only at the end of the experiment. If we adopt the standard methodology, the average span is 3.08 for our school age students and 4.16 for our undergraduates. These numbers are comparable to existing studies reporting averages between 3 and 4 in children and adolescents (Hale et al., 2002) and between 4 and 5 in adults (Gignac, 2015).

	LILA						USC
Grade	6th	7th	8th	9th	10th	11th	U
Correct	40.2	45.3	45.6	47.0	52.5	54.8	70.1

Table 4: Average percentage of correct trials in Digit Span Task by grade

in itself, we are more interested in correlating it with behavior in our main task. Optimal performance in the treasure game necessitates the combination of several abilities. First, the individual must realize that offers other than v^+ (in o) and v_i^+ (in h and l) are necessarily suboptimal. Then, the subject must be able to determine which offer among $\underline{v}^+$ and $\overline{v}^+$ maximizes payoffs depending on the round. To find out which abilities correlate with working memory ability, we perform OLS regressions at the individual level, where the dependent variable is performance in the Digit Span Task. We consider only school-age participants and use as regressors *Age* in months and the same control variables as before (gender, siblings and favorite topic). We also include a combination of performance in o and l (column (3)) or performance in h and l (column (4)) to measure the contribution of the different cognitive abilities. Notice that we do not include performance in o and h in the same regressions since we know from Table 2 that these two measures are highly colinear. The results are reported in Table 5.

Confirming the existing results in the literature (Gathercole et al., 2004), performance in the Digit Span Task increases significantly with age. Individuals with an inclination for analytical topics (self-reported preference for STEM) perform better. Most notably, there is a positive association between optimal choice and working memory. This is particularly visible for choice in the simpler o- and h-rounds, which both very significantly predict performance in the Digit Span Task. The positive but statistically less significant effect of performance in l-rounds suggest that working memory also supports, although to a lesser extent, the more subtle step that consists in discriminating between $\underline{v}^+$ and $\overline{v}^+$.

Given that the (potential) link between working memory and contingent reasoning is correlational, our next set of regressions uses performance in the Digit Span Task as an explanatory variable of the participant's type in the treasure game. Participants classified as **O** have the lowest scores (4.35) while those classified as **R** have the highest scores (6.33). The score of **O** types is significantly lower than that of **H**, **S** and **R** and the differences between **H** and **R** are also significant ($p < 0.04$ after correcting for multiple comparisons). Naturally, these effects are partly driven by age differences across types. To control for age effects, we present in Table 6 a multinomial logistic regression of the type of each participant similar to the one reported in Table 3. However, we group types **L**, **H** and **S** in one category to increase statistical power (the new default), and we add

	(1)	(2)	(3)	(4)
Age	0.026***	0.026***	0.023***	0.021**
	(0.007)	(0.007)	(0.007)	(0.007)
STEM	—	0.718**	0.592*	0.624*
		(0.273)	(0.271)	(0.270)
Male	—	-0.385	-0.382	-0.421
		(0.264)	(0.265)	(0.263)
Siblings	—	-0.590	-0.596	-0.582
		(0.325)	(0.319)	(0.319)
correct-o	—	—	0.403**	—
			(0.155)	
correct-h	—	—	—	0.197**
				(0.072)
correct-l	—	—	0.066	0.110*
			(0.058)	(0.055)
Constant	1.362	1.606	1.031	1.316
	(1.104)	(1.097)	(1.143)	(1.111)
Adj. R^2	0.054	0.094	0.111	0.114
# obs.	261	261	261	261

(standard errors in parenthesis)
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 5: OLS regressions of individual performance in the Digit Span Task as a function of demographic variables and performance in the treasure game

the performance in the Digit Span Task as an independent variable (*DigitSpan*).

Table 6 reinforces the results from Table 5. After controlling for age and other individual characteristics, we still observe that performance in the working memory task is lower by **O** subjects and higher by **R** subjects. The result indicates that a better working memory helps performance significantly, especially in the simpler rounds. At the same time, it is not sufficient for optimal behavior, and other cognitive processes are also involved in the calculations required by contingent reasoning. This finding has implications for the study of complex games that require a priori sophisticated logical skills.

Last, note that while we are primarily interested in the developmental trajectory of performance in contingent reasoning and how working memory affects it, we have observed that gender and topic preferences are also often significantly associated with the main measures. These effects may be overestimated due to measurement errors and imperfect correlations between explanatory variables. We address these issues in Appendix C and confirm the significant association between gender and complex contingent reasoning, as well as the association between topic preference and working memory.

	O	**R**
Age	-0.001	0.028***
	(0.010)	(0.008)
STEM	-0.172	0.359*
	(0.420)	(0.318)
Male	0.036	0.953***
	(0.388)	(0.315)
Siblings	-0.477	-0.199
	(0.458)	(0.387)
DigitSpan	-0.288***	0.124*
	(0.101)	(0.074)
Constant	0.564	-6.620***
	(1.645)	(1.393)
AIC	467.466	467.466
# obs.	261	261

(standard errors in parenthesis)
* p < 0.05; ** p < 0.01; *** p < 0.001

Table 6: Multinomial logistic regression of types on demographic variables and performance in the Digit Span task (defaults = **L**, **H** and **S**)

5 Conclusion

Our study shows that performance in deterministic problems involving contingent reasoning is critically linked to the complexity of the task. The most basic aspects are grasped even by our youngest participants but the ability to solve the most subtle aspects develops gradually with age. It is not facilitated by repeated exposure or feedback. A higher score in working memory is positively associated with performance.

The study approaches contingent reasoning from a developmental perspective. It brings further evidence that age plays a critical role in the development of strategic thinking. Previous studies have shown that children develop inductive logic between the ages of 8 and 12 (Feeney and Heit, 2007) and the ability to perform hypothetical and counterfactual thinking between the ages of 11 and 14 (Piaget, 1972; Rafetseder et al., 2013; De Neys and Everaerts, 2008). Observing that conditional reasoning abilities develop during that time period and beyond is consistent with that evidence. On the other hand, we have recently reported a steep development throughout elementary school (ages 6 to 11) and a lack of improvement past that period in backward induction games (Brocas and Carrillo, 2021b). Stagnation occurs even though the ability is not necessarily mastered at that age, indicating that children may have reached a cognitive bound. In other research on games of strategy, we have also documented no improvement with age in optimal randomization

in hide and seek games (Brocas and Carrillo, forthcoming). The significant improvement in our treasure game during middle and high school contrasts with that evidence, and implies that developmental trends and cognitive boundaries are situation-specific. It also suggests that age effects may extend beyond adolescence into young adulthood.

Despite the caveats noted, the results obtained here are consistent with [MNV]. Our oldest school-age participants and undergraduates perform better than in the traditional 'acquire-a-company' game. There is no evidence of cursedness, indicating that this behavioral feature may be tightly linked to uncertainty. The increased performance of our subjects relative to the deterministic game in [MNV] is likely due to our higher control level and more educated population (lab-in-the field experiment with a homogenous population of private school high schoolers and college undergraduates vs. MTurk workers) as well as the simpler, graphical design. Finally, the responses of participants support the idea that uncertainty is an added obstacle to optimal performance.

Our study helps clarify the contribution of several components of contingent reasoning in the context of adverse selection games. [MNV] have argued that contingent reasoning can be decomposed into uncertainty and computational complexity. By removing uncertainty, the problem is simplified and performance depends on the ability to make complex calculations. We take a different stand and argue instead that the problem deprived from uncertainty incorporates the same qualitative reasoning difficulties as the original problem. It can be decomposed into different logical requirements: conditional thinking ("if a seller is of a certain type, they will accept that set of offers"), recursive logic ("given the set of offers acceptable by the seller, what is the optimal price to offer?"), counterfactual thinking ("what about a seller with a different type?") and further recursive thinking to combine the outputs of the previous pieces into a comparison of type-dependent optimal strategies. Our results show that difficulties relate to recursive and counterfactual logic, not to conditioning per se. Algebraic or other basic computational difficulties also seem to have a limited effect. Furthermore, since recursive and counterfactual thinking abilities improve during adolescence and they heavily tax working memory (which develops at the same time), we observe that performance progresses gradually over that period. Overall, working memory turns out to be an essential contributor to the recursive and counterfactual thinking abilities involved in adverse selection games.

The data reveals an unexpected gender difference. Conditional reasoning is a logical ability and there is no evidence of a specific gender bias in the literature studying logic, cognition and IQ. At best, results are mixed and differences are typically small (Lynn and Irwing, 2004; Reynolds et al., 2008). Still, gender differences have been observed in game theoretic settings involving steps of reasoning in adults (Cubel and Sanchez-Pages, 2017) and children (Brocas and Carrillo, 2021b). Depending on games and treatments,

males or females may perform better. It is, however, intriguing and worthy of further investigation that the same gender effect is present in our study and in [MNV], despite the very significant differences in the populations studied.

Participants with a preference for science tend to have a higher performance in the working memory task, which in turn affects performance in the contingent reasoning task. This is consistent with the literature showing that working memory is associated with academic achievement in science-related topics (Alloway and Alloway, 2010; Swanson, 2011) and with studies showing a relationship between high cognition and performance in games (Brañas-Garza et al., 2012; Gill and Prowse, 2016; Proto et al., 2019, 2020; Fe et al., 2020). It is also in line with our earlier research where we showed that a preference for science is associated with better performance and higher payoffs (Brocas and Carrillo, 2021b). The present study also suggests that a highly developed working memory capacity is a necessary condition to solve complex reasoning tasks, although it may not be sufficient.

Our results also illustrate the methodological value of studying basic cognitive abilities and assessing their relationship with performance in games. There is converging evidence that cognition and decision-making go hand-in-hand. Cognitive development and age-related changes in decision-making cannot be dissociated. Basic cognitive functions refer to simple processes such as working memory (the ability to manipulate pieces of information) and inhibitory control (the ability to reject distracting or irrelevant features). These functions are recruited to complete simple cognitive tasks (such as number manipulation and mental rotation of objects) but also complex cognitive tasks (such as contingent reasoning and deductive logic). Decision-making is a form of complex cognitive activity, which often combines several kinds of logic to be carried efficiently. Therefore, non-compliance to central predictions of theory may be traced to several fundamentally different causes. Testing cognitive abilities independently allows us to identify which feature of the decision poses a problem.[30] In our case, low performance in working memory accounts for a large part of suboptimal choices. It suggests that it is important to link non-equilibrium play to limitations of higher-level cognitive skills (e.g., working memory, cognitive flexibility, inhibitory control) that control and coordinate lower-level logical abilities (e.g., logical deductions) and behavior. We believe that a better understanding of the relative contribution of each high-level skill has long range implications for the design of behavioral models. Also, while studying the relationship between basic cognitive abilities and decision-making is particularly interesting in children, it should also be enlightening

[30] There exists a large variety of tasks that assess cognitive functions. Some are specific, such as the Digit Span task and its homologue spatial form the Corsi task. Their objective is to diagnose a well-defined ability. Others are general, such as IQ tests. Their objective is to provide a score regarding the overall cognitive level of a person.

in studies with adults.

The results have vast implications for policy intervention. Indeed, contingent reasoning is ubiquitous in strategic interactions. Developing this ability is critical to assess the motivations of others and to anticipate their reactions to our choices. A person who does not apply contingent reasoning properly can be taken advantage of by people who misrepresent their intentions. This is particularly problematic in the case of adolescents, a population often targeted by cyber threats and scams. Our results–which demonstrate that contingent reasoning is developing during adolescence–suggest that intervention is needed to protect teens in situations they are not equipped to handle.

References

George A Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488–500, 1970.

S Nageeb Ali, Maximilian Mihm, Lucas Siga, and Chloe Tergiman. Adverse and advantageous selection in the laboratory. *American Economic Review*, 111(7):2152–78, 2021.

Tracy Packiam Alloway and Ross G Alloway. Investigating the predictive roles of working memory and iq in academic attainment. *Journal of experimental child psychology*, 106 (1):20–29, 2010.

Alan D Baddeley and Graham Hitch. Working memory. In *Psychology of learning and motivation*, volume 8, pages 47–89. Elsevier, 1974.

Sheryl B Ball, Max H Bazerman, and John S Carroll. An evaluation of learning in the bilateral winner's curse. *Organizational Behavior and human decision Processes*, 48(1): 1–22, 1991.

Yoella Bereby-Meyer and Brit Grosskopf. Overcoming the winner's curse: an adaptive learning perspective. *Journal of Behavioral Decision Making*, 21(1):15–27, 2008.

Karin Binder, Stefan Krauss, and Georg Bruckmaier. Effects of visualizing statistical information–an empirical study on tree diagrams and 2× 2 tables. *Frontiers in psychology*, 6:1186, 2015.

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.

Gary L Brase. The power of representation and interpretation: doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. *Journal of Cognitive Psychology*, 26(1):81–97, 2014.

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*, 179:777–783, 2020a.

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, 2020b.

Isabelle Brocas and Juan D Carrillo. Dynamic coordination in efficient and fair strategies: a developmental perspective. USC Working Paper, 2021a.

Isabelle Brocas and Juan D Carrillo. Steps of reasoning in children and adolescents. *Journal of Political Economy*, 129(7):2067–2111, 2021b.

Isabelle Brocas and Juan D Carrillo. The development of randomization and deceptive behavior in mixed strategy games. *Quantitative Economics*, forthcoming.

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

Juan D Carrillo and Thomas R Palfrey. No trade. *Games and Economic Behavior*, 71(1): 66–87, 2011.

Marco Casari, Jingjing Zhang, and Christine Jackson. Same process, different outcomes: group performance in an acquiring a company experiment. *Experimental Economics*, 19 (4):764–791, 2016.

Gary Charness and Dan Levin. The origin of the winner's curse: a laboratory study. *American Economic Journal: Microeconomics*, 1(1):207–36, 2009.

Gary Charness, John A List, Aldo Rustichini, Anya Samek, and Jeroen Van De Ven. Theory of mind among disadvantaged children: Evidence from a field experiment. *Journal of Economic Behavior & Organization*, 166:174–194, 2019.

Daniel L Chen, Martin Schonger, and Chris Wickens. otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, 2016a.

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, 2016b.

Ramon Cobo-Reyes, Jose J Dominguez, Fernando García-Quero, Brit Grosskopf, Juan A Lacomba, Francisco Lagos, Tracy Xiao Liu, and Graeme Pearce. The development of social preferences. *Journal of Economic Behavior & Organization*, 179:653–666, 2020.

David J Cooper and Matthias Sutter. Endogenous role assignment and team performance. *International Economic Review*, 59(3):1547–1569, 2018.

Vincent P Crawford and Nagore Iriberri. Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? *Econometrica*, 75(6):1721–1770, 2007.

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*, 2016.

Wim De Neys and Deborah Everaerts. Developmental trends in everyday conditional reasoning: The retrieval and inhibition interplay. *Journal of Experimental Child Psychology*, 100(4):252–263, 2008.

Iroise Dumontheil. Development of abstract thinking during childhood and adolescence: The role of rostrolateral prefrontal cortex. *Developmental cognitive neuroscience*, 10: 57–76, 2014.

Randall W Engle, Stephen W Tuholski, James E Laughlin, and Andrew RA Conway. Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of experimental psychology: General*, 128(3):309–331, 1999.

Ignacio Esponda. Behavioral equilibrium in economies with adverse selection. *American Economic Review*, 98(4):1269–91, 2008.

Erik Eyster and Matthew Rabin. Cursed equilibrium. *Econometrica*, 73(5):1623–1672, 2005.

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

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

Susan E Gathercole, Susan J Pickering, Benjamin Ambridge, and Hannah Wearing. The structure of working memory from 4 to 15 years of age. *Developmental psychology*, 40 (2):177, 2004.

Gilles E Gignac. The magical numbers 7 and 4 are resistant to the flynn effect: No evidence for increases in forward or backward recall across 85 years of data. *Intelligence*, 48:85–95, 2015.

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.

Ben Gillen, Erik Snowberg, and Leeat Yariv. Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy*, 127(4):1826–1863, 2019.

Brit Grosskopf, Yoella Bereby-Meyer, and Max Bazerman. On the robustness of the winner's curse phenomenon. *Theory and Decision*, 63(4):389–418, 2007.

James B Hale, Jo-Ann B Hoeppner, and Catherine A Fiorello. Analyzing digit span components for assessment of attention processes. *Journal of Psychoeducational Assessment*, 20(2):128–143, 2002.

Simon J Handley, A Capon, M Beveridge, I Dennis, and J St BT Evans. : Working memory, inhibitory control and the development of children's reasoning. *Thinking & Reasoning*, 10(2):175–195, 2004.

William T Harbaugh and Kate Krause. Children's altruism in public good and dictator experiments. *Economic Inquiry*, 38(1):95–109, 2000.

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.

Arthur R Jensen and Richard A Figueroa. Forward and backward digit span interaction with race and iq: Predictions from jensen's theory. *Journal of Educational Psychology*, 67(6):882, 1975.

John H Kagel and Dan Levin. The winner's curse and public information in common value auctions. *The American economic review*, pages 894–920, 1986.

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.

Alejandro Martínez-Marquina, Muriel Niederle, and Emanuel Vespa. Failures in contingent reasoning: The role of uncertainty. *American Economic Review*, 109(10):3437–74, 2019.

Michelle McDowell and Perke Jacobs. Meta-analysis of the effect of natural frequencies on bayesian reasoning. *Psychological bulletin*, 143(12):1273, 2017.

J Keith Murnighan and Michael Scott Saxon. Ultimatum bargaining by children and adults. *Journal of Economic Psychology*, 19(4):415–445, 1998.

Klaus Oberauer, H-M Süß, Ralf Schulze, Oliver Wilhelm, and Werner W Wittmann. Working memory capacity—facets of a cognitive ability construct. *Personality and individual differences*, 29(6):1017–1045, 2000.

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*, 127(3):1351–1390, 2019.

Eugenio Proto, Aldo Rustichini, and Andis Sofianos. Intelligence, errors and strategic choices in the repeated prisoners' dilemma. 2020.

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.

Brian W Rogers, Thomas R Palfrey, and Colin F Camerer. Heterogeneous quantal response equilibrium and cognitive hierarchies. *Journal of Economic Theory*, 144(4):1440–1467, 2009.

W Samuelson and MH Bazerman. The winner's curse in bilateral negotiations. vl smith, ed. research in experimental economics, vol. 3. *JAI Press, Greenwich, CT*, 105:137, 1985.

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.

Doron Sonsino, Ido Erev, and Sharon Gilat. On rationality, learning and zero-sum betting–an experimental study of the no-betting conjecture. Working Paper, 2002.

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.

H Lee Swanson. Working memory, attention, and mathematical problem solving: A longitudinal study of elementary school children. *Journal of Educational Psychology*, 103 (4):821, 2011.

Joseph V Terza, Anirban Basu, and Paul J Rathouz. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of health economics*, 27(3):531–543, 2008.

David Wechsler. Wechsler intelligence scale for children. 1949.

Appendix A. Experimental details

A1. Instructions - LILA

Today, we are going to play a few games with you. In those games, you will earn points that will be converted into money and placed in your virtual wallet. Each point is worth 2 cents. At the end of the session, we will send to your LILA email address an Amazon gift card with the money you earned. You will start with 300 points, so $6.00.

Treasure Game

In this game you will play against robots. Robots are computer programs that play in a predetermined way. We are going to tell you how. Here is an example. This is the computer you are playing with [point] and this is you [point] (see Figure 7 for the slides).

[SLIDE 1: screen with one robot]

The computer owns a treasure box. The box has two compartments, an upper compartment and a lower compartment. There are points in each compartment [point]. However, the computer does not have the key of the lower compartment and cannot access the points placed there. So, even though the computer knows how many points there are in the lower compartment, the box for the computer is worth only the number of points in the top compartment.

Now, you do have the key of the lower compartment. So, if you buy the treasure box, you will be able to access it. This means that the box is worth to you the points in the upper and lower compartments.

To buy a box, you need to make an offer to the computer, and the offer has to be accepted. Remember that, for the computer, the box is worth only the number of points in the top compartment. If the offer it receives is equal or above that number, it will accept the deal. Otherwise, there will be no deal. The computer is programmed, so this is an automatic rule.

Each time you make an offer, you will know if the trade occurs and how much money you win or lose. Let's look at what can happen.

[SLIDE 2: screen with one robot and DEAL]

If you offer 60 (and this is just an example), the computer will accept it and you will be notified that there is a deal. You will also learn how many points you earned. In this case, $70 - 60 = 10$ points.

[SLIDE 3: screen with one robot and NO DEAL]

But, if you offer 40 (and, again, this is just an example), the computer will not accept it and you will be notified that the there is no deal. In that case, you will earn 0. To sum up, if your offer is not accepted, you get 0. If your offer is accepted, you accumulate points on your wallet.

Now, there are two types of trading games. At the beginning, you will play with one computer, as we have just described. After a few rounds, you will play with two computers. In that case, you will see a screen like this.

[SLIDE 4: screen with two robots]

This is you and these are the two computers you are playing with. What is important to realize is that you have to make **the same offer** to both computers. Each computer sees the offer

and decides whether to accept the deal or not. As before, each computer accepts your offer if it is equal or greater to the number of points in the upper compartment. Therefore, it may be that no computer accepts your offer, one computer accepts your offer and the other doesn't or both computers accept your offer. This also means that you may win points with both computers, or you may lose with some computer. If this happens, don't worry. The points we give you in advance will cover your losses. Finally, your offer can never exceed the total value of the most valuable box. In this example, you cannot offer more than 70.

Practically speaking, you need to enter your offer in one cell and the other will automatically be populated. You are going to play several rounds. Each time, the boxes will have different values. Is that clear?

Before playing, you will answer a short quiz. This is not a test. We are just trying to make sure you have understood the rules because it is important to understand the rules when you play a game. All the questions in the quiz refer to this screen.

[SLIDE 5: screenshot for quiz]

[Launch quiz] - [When quiz is done] OK, now we will launch the game. You will be playing several times [Launch game]

The game is over, but we would like you to answer two questions about the game. Read carefully. In the first question, you will see a screenshot like the one in the game you have been playing. In the second question, you will see a screenshot slightly different. The rules are the same as before except that you are playing against one of the two computers, but you do not know which one (each of them is equally likely) [Launch questions]

Digit memory game

In this game, you will see digits, one at a time, on your screen. Each digit will appear for half a second. Then, a new digit will appear after a little bit less than a second. You need to pay attention to the digits and the order in which they are shown because you will have to report them in the reverse order and press OK. Here is an example:

[SLIDE 6: launch video 1]

What were the digits? 4 - 8 - 3. Good, so you need to enter them in reverse order: 3 - 8 - 4, and press OK as in the following video.

[SLIDE 7: launch video 2]

You will start with 3 digits and keep increasing until you get to 8 digits. Each time you enter the correct answer, you will earn 20 points and each time you don't enter the correct answer, you will get 0 points, so pay attention! As in the video, the cursor will be automatically set in the left box. Once you input a digit, you will not be able to change it, and the cursor will automatically move to the next box. When you are done, press OK to move to the next set of numbers. Any questions? Please be very quiet during this game [Launch game]

We are done. You will now see the number of points you got in the Treasure game and in the Digit Memory game. You don't need to memorize them. Press OK and fill the questionnaire while we prepare your payment [Launch final questionnaire].

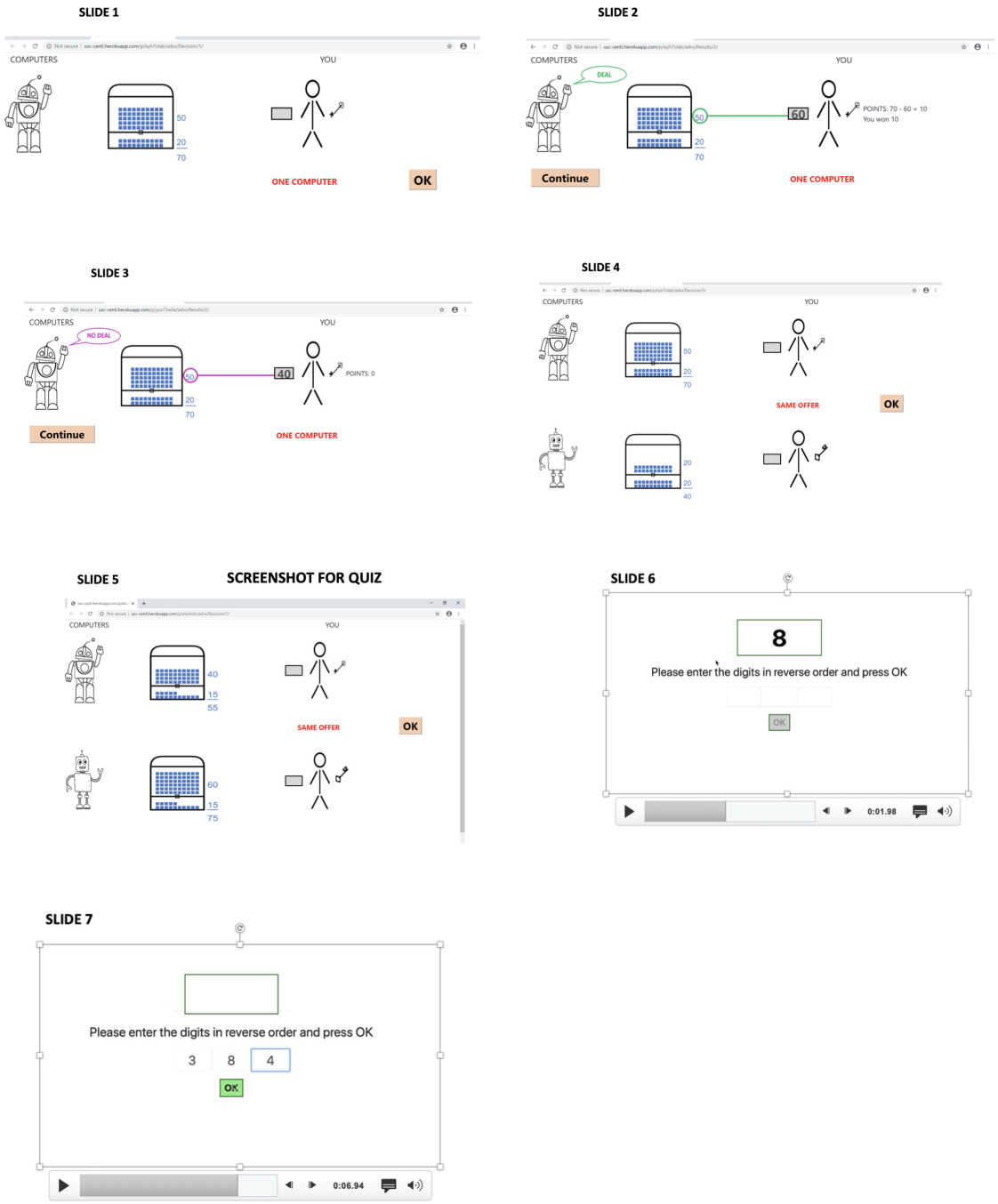


Figure 7: Slides used for the instructions

A2. Quiz

[Questions refer to the screenshot projected in SLIDE 5. Correct answers marked in **bold**].

Suppose you observe this screen, answers the following questions.

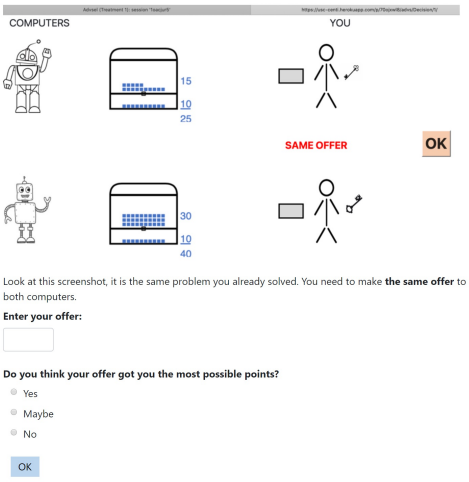
1. How many offers can you make?
   - **a. You have to make the same offer to both computers**
   - b. You can make different offers to each computer
   - c. It doesn't say

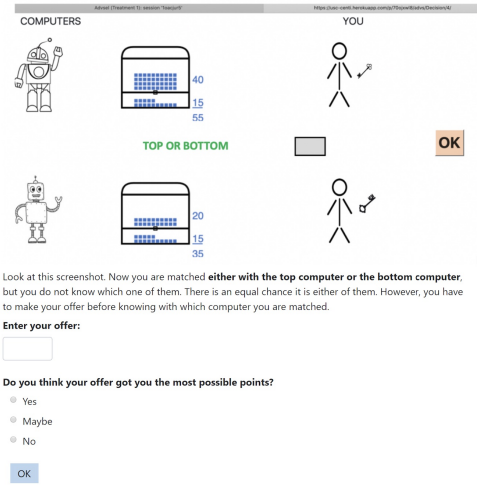
2. Suppose you offer 65 to each computer. What happens?
   - a. No Deal with top computer and No Deal with bottom computer
   - b. No Deal with top computer and Deal with bottom computer
   - c. Deal with top computer and No Deal with bottom computer
   - **d. Deal with top computer and Deal with bottom computer**

3. Suppose you offer 50 to each computer. How many points do you get in your relationship with the top computer (for this exercise do not count what happens with the bottom computer)
   - **a. There is a deal and you win 5 points**
   - b. There is a deal and you win 10 points
   - c. There is a deal and you lose 5 points
   - d. There is no deal so you get 0 points

4. Suppose you offer 50 to each computer. How many points do you get in your relationship with the bottom computer (for this exercise do not count what happens with the top computer)
   - a. There is a deal and you win 5 points
   - b. There is a deal and you win 10 points
   - c. There is a deal and you lose 5 points
   - **d. There is no deal so you get 0 points**

A3. Two-part, non-incentivized questionnaire





A4. Payoff combinations in the treasure game

Table 7 reports the payoffs in each chest for each of the 15 rounds. Values in **bold** reflect the optimal offer in that round.

Round		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Chest 1	upper	**45**	**15**	**30**	20	**40**	60	40	**25**	**50**	**30**	45	**40**	20	30	**30**
	lower	10	20	15	10	15	20	15	15	20	10	15	15	20	15	8
Chest 2	upper	–	–	–	**25**	30	**30**	**15**	15	40	45	**20**	30	**35**	**10**	40
	lower	–	–	–	10	15	20	15	15	20	10	15	15	20	15	8

Table 7: Payoffs by compartment and chest

Appendix B. Summary of behavior in other populations

Table 8 provides a descriptive summary of the behavior by the 8 math teachers at LILA (Teachers) and the 11 Master students at USC (Masters).

	Optimal play			Type				
	o-round	*h*-round	*l*-round	**R**	**H**	**L**	**S**	**O**
Teachers	1	0.9	0.58	0.25	0.25	0.00	0.50	0.00
Masters	0.79	0.76	0.68	0.55	0.18	0.09	0.09	0.09

	Correct question				Confidence Det.			Confidence Prob.		
	BOTH	DET	PROB	NONE	Yes	Maybe	No	Yes	Maybe	No
Teachers	0.75	0.125	0.00	0.125	0.62	0.38	0.00	0.25	0.75	0.00
Masters	0.73	0.00	0.00	0.27	0.91	0.09	0.00	0.55	0.36	0.09

Table 8: Summary statistics of behavior by Teachers (LILA) and Masters (USC)

The samples are extremely small to make meaningful inferences. Also, these two populations are relatively heterogenous (in particular, USC master students come from very different intellectual backgrounds). The results, however, indicate that Teachers and Masters do not behave very differently from 11th grades and undergraduates.

Appendix C. Robustness

Two potential issues may bias the coefficients obtained in the regression analysis of section 4.2. First, some of the exogenous variables are correlated to some extent. Second, our behavioral measures, the main variables of interest, are subject to measurement errors.

A fix for the latter problem is to duplicate measures and use one to instrument the other (Gillen et al., 2019). Because we have not designed our experiment to allow for that possibility, we cannot use that specific method. However we can provide a few analyses designed to better account for the two issues above.

The OLS regression in Table 5 reports a consistent significant effect of the control variable *STEM* on the endogenous variable. In those regressions *Age*, correct-o, correct-h and correct-l are correlated to some extent, and the last three also suffer from measurement errors. A simple way to better assess the effect of *STEM* is to run a principal component analysis of the four imperfectly correlated variables, then run a regression on the principal components and control variables. Coefficients on the principal components are not readily interpretable, but the method ensures that the regressors are independent. For our case, the first principal component (*PC1*) explains 44% of the variance, the second (*PC2*) 28%, and the third (*PC3*) 19%. We retain these three for the regression analysis. Table 9 shows that *STEM* continues to be associated with performance in the Digit Span Task, while *Male* and *Siblings* continue to not be associated with performance in the Digit Span Task.

	(1)	(2)	(3)
STEM	0.663*	0.618*	0.602*
	(0.270)	(0.269)	(0.270)
Male	-0.414	-0.362	-0.393
	(0.261)	(0.260)	(0.263)
Siblings	-0.609	-0.595	-0.589
	(0.320)	(0.317)	(0.318)
PC1	-0.091***	-0.237***	-0.246***
	(0.019)	(0.065)	(0.066)
PC2	—	0.082*	0.055
		(0.035)	(0.050)
PC3	—	—	-0.031
			(0.042)
Constant	0.791	0.826	1.062
	(1.078)	(1.069)	(1.116)
Adj. R^2	0.105	0.120	0.118
# obs.	261	261	261

(standard errors in parenthesis)
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 9: OLS regressions of individual performance in the Digit Span Task as a function of demographic variables and principal components

Variables *Age* and *DigitSpan* are also imperfectly correlated in the multinomial logistic

regression in Table 6. We can therefore perform a similar principal component analysis. Table 10 presents the alternative multinomial regression with one principal component (*PC*, which explains 62% of the variance) replacing the variables *Age* and *DigitSpan* in Table 6. As in the previous regression, males are still more likely to be classified as rational. On the other hand, the effect of STEM is no longer significant. A similar result is achieved with an instrumental variable approach, whereby one first regresses *DigitSpan* on demographic variables (first step) and then uses the predicted *DigitSpan* value in the multinomial logistic regression (second step).[31]

	O	**R**
STEM	-0.290	0.427
	(0.413)	(0.312)
Male	0.248	0.933***
	(0.373)	(0.313)
Siblings	-0.242	-0.258
	(0.440)	(0.383)
PC	0.015***	-0.043***
	(0.014)	(0.011)
Constant	0.506	-6.619***
	(1.641)	(1.378)
AIC	475.580	475.580
# obs.	261	261

(standard errors in parenthesis)
* p < 0.05; ** p < 0.01; *** p < 0.001

Table 10: Multinomial logistic regression of types on demographic variables and principal component (defaults = **L**, **H** and **S**)

Overall, the results suggest that gender and topic preferences influence task performance differentially. While gender is consistently associated with behavior in the treasure game (Tables 2, 3, 6 and 10), topic preference seems to impact primarily performance in the Digit Span Task (Tables 5 and 9), and influences behavior only indirectly.

[31] In our case, the instrument is weak. We ran the procedure including the residuals from the first step into the second regression as suggested by Terza et al. (2008).