The development of consistent decision-making across economic domains *

Isabelle Brocas
University of Southern California and CEPR

Juan D. Carrillo
University of Southern California and CEPR

T. Dalton Combs †

Niree Kodaverdian †

April 2019

Abstract

How does value-based reasoning develop and how different this development is from one domain to another? Children from Kindergarten to 5th grade make pairwise choices in the Goods domain (toys), Social domain (sharing between self and other), and Risk domain (lotteries) and we evaluate how consistent their choices are. We report evidence that the development of consistency across domains cannot be fully accounted for by existing developmental paradigms: it is not associated to the development of transitive reasoning and it is only partially linked to the age-related changes in attentional control and in the tendency to focus on a subset of choice attributes (centration). We show that choice consistency is related to self knowledge of preferences which develops *gradually* and *differentially* across domains. The Goods domain offers a developmental template: children become more consistent over time because they learn what they like most and least. In the Social domain, children gradually learn what they like most but not what they like least, while in the Risk domain, they gradually learn what they like least but not what they like most. These asymmetric developments give rise to asymmetric patterns of consistency.

Keywords: laboratory experiments, developmental economics, revealed preferences, risk, social preferences.

JEL Classification: C91, D11, D12.

*We are grateful to members of the Los Angeles Behavioral Economics Laboratory (LABEL) for their insights and comments in the various phases of the project. We thank the editor of this journal, two anonymous referees, and the participants at the 2014 Social Neuroscience retreat (Catalina Island, USC) and at the 2015 Morality, Incentives and Unethical Behavior Conference (UCSD) for useful comments. We also thank Nordine Bouriche and all the staff at the Lycée International de Los Angeles for their help and support. All remaining errors are ours. 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-1425062.

†These authors contributed to the first version of the manuscript while they were PhD students at the University of Southern California.

1 Introduction

Adults have many decision-making abilities that children lack. An important question is how these abilities develop and what promotes them. In this study, we investigate the developmental trajectory of the most fundamental ability for economic decision-making, that of making consistent choices. We measure consistency by testing for transitivity of choices: if a subject chooses option A over option B and option B over option C, consistency requires that she chooses option A over option C. A number of studies have shown that children, especially young ones, are less consistent than adults when choosing between foods or toys, from now referred to as the "Goods" domain (Smedslund, 1960; Harbaugh, Krause, and Berry, 2001; List and Millimet, 2008). Other studies suggest that age is less of a predictor of consistency in situations involving sharing between oneself and another person, from now referred to as the "Social" domain (Harbaugh and Krause, 2000). Finally the evidence is also mixed in options involving lotteries, from now referred to as the "Risk" domain (Harbaugh, Krause, and Vesterlund, 2002).[1]

Multiple paradigms may explain why differences exist between adult and children behavior and why they change during development. Even if children know that they prefer A to B, their choices may not conform to this ranking for reasons related to established developmental paradigms. For example, it could occur because attentional control, a capacity which has previously been associated with transitive behavior in older adults (Brocas, Carrillo, Combs, and Kodaverdian, 2019), is still underdeveloped in children (Davidson, Amso, Anderson, and Diamond, 2006; Astle and Scerif, 2008). Alternatively, it may be because the ability to reason logically (Sher, Koenig, and Rustichini, 2014; Barash, Brocas, Carrillo, and Kodaverdian, 2019; Brocas and Carrillo, 2018b) and transitively (Piaget, 1948; Bouwmeester and Sijtsma, 2006) is still developing. Finally, it may result from children's inability to focus on more than one attribute of an item at a time, a phenomenon referred to as centration (Piaget, Elkind, and Tenzer, 1967; Donaldson, 1982; Crain, 2015).

The objective of this research is to assess the *common* and *domain-specific* developmental trajectories of transitive decision-making in the Goods, Social, and Risk domains in children from Kindergarten to 5th grade and to determine if the dominant developmental paradigms (attentional control, logical reasoning, and centration) are enough to explain age- and domain-related differences in transitive decision-making. Assessing consistency across domains is critical to the understanding of economic choice because it allows us to identify domain-specific features that enhance or prevent consistency. Relating consistency to underlying abilities is essential to describe the mechanisms that promote quality

[1] For a survey on economic experiments that test behavior of children and, in particular, their ability to make rational decisions, see Sutter, Zoller, and Glätzle-Rützler (2019).

of decision-making.

We therefore designed a novel transitivity paradigm to obtain comparable measurements of consistent decision-making across the three above mentioned domains: Goods, Social, and Risk. For each domain, we asked participants to make pairwise decisions between options and we evaluated how transitive their decisions were. We recorded the number of transitivity violations and we assessed the severity of those violations. For analysis, we also asked participants to provide an explicit ranking of these options, which could then be compared to their choices. We added trivial trials to assess attention and we included a transitive reasoning task to probe logical reasoning. Finally, we identified centration tendencies by categorizing participants whose decisions were compatible with choices based on the evaluation of one attribute at a time. We also compared the results obtained in children with adult-level performance.

Our study is most closely related to the literature on choice consistency. There are however two main differences between the existing studies and ours. A first difference is conceptual. We want to compare consistency across domains. Earlier studies offer instead analyses of consistency in one domain at a time. A second difference is methodological. Earlier studies have relied on the Generalized Axiom of Revealed Preferences (GARP), an *indirect* test of transitivity which focuses on choices between bundles of options given a budget constraint, a system of prices, and a non-satiation assumption (Samuelson (1938), Varian (1982), and others). By contrast, our design is a *direct* test of transitivity. The extreme simplicity of the design (pairwise comparisons instead of a system of prices and a budget constraint) and the removal of assumptions such as non-satiation (sometimes violated in experimental studies) makes it especially suitable for testing choices in a population of non-highly educated individuals, such as our young children. It also delivers results that, unlike GARP tests, are comparable across domains. Overall, our novel design provides a unique tool to address the question of consistency across domains and ages.

Our study of consistency jointly tests completeness and transitivity of preferences. We conjecture that the preferences of our subjects are complete, because we did not notice any confusion, delay or difficulty in making choices in any of our tasks. However, the design of our study is such that we cannot rule out failure of completeness as an alternative explanation of our results.

We made four critical hypotheses. First and consistent with the existing literature, we conjectured that in all domains the number of transitivity violations would be high for the youngest children and would monotonically decrease with age to reach a level close to zero in the adult population. Second, we hypothesized that the differences across domains observed in the literature exist because designs –hence results– are not directly comparable. Within our design, the number of transitivity violations should be similar

across domains for each age group. Third, it is intuitive that some options are less prone to be involved in violations, because they are highly desirable or, on the contrary, very unattractive. Such options are easy to select or avoid consistently. We therefore theorized that the most and least preferred items would be less often involved in transitivity violations. Last, given that development during elementary school is known to revolve around changes in centration, attention and (simple) logical reasoning, we hypothesized that the development of choice consistency across domains would be fully explained by these changes. Overall, we anticipated a template in which inconsistencies mostly involve options for which participants did not have strong preferences for or against, and progressively decrease with age in conjunction with an increase in attention and logical reasoning abilities as well as changes in centration tendencies. These four hypothesis were tested separately to answer two central questions: is the developmental trajectory of consistent decision-making domain-specific and what drives the trajectory?

The results obtained for participants in the adult group were consistent with the typical findings reported previously in the literature, according to which, by adulthood, people are largely consistent in the Goods (Battalio, Kagel, Winkler, Fisher, Basmann, and Krasner, 1973; Cox, 1997) and Social (Andreoni and Miller, 2002; Fisman, Kariv, and Markovits, 2007) domains. Our findings in the Risk domain agreed with the original transitivity study (Loomes, Starmer, and Sugden, 1991), but it was less optimistic than recent GARP studies (Mattei, 2000; Choi, Fisman, Gale, and Kariv, 2007). The trajectories towards these adult outcomes were different across domains. Behavior in the Goods domain closely followed the expected template, with significant improvements with age, high consistency among 4th-5th graders and more mistakes for options ranked in the intermediate range. In the Social domain, many young participants utilized rules of behavior compatible with centration. Interestingly, these rules made them look relatively more consistent in our paradigm. Among children who did not use such simple rules, we observed the same trajectory as in the Goods domain, suggesting that centration concealed their underdeveloped decision-making abilities. In the Risk domain, participants were not nearly as consistent as in the other two domains. Some (young) children exhibited centration tendencies in their choices. However, among those who did not, all performed at the same level, suggesting that the ability to trade-off probabilities and rewards was not yet developed, perhaps partly as a consequence of a still underdeveloped working memory system (Gathercole, Pickering, Ambridge, and Wearing, 2004). Interestingly, performance was correlated across domains among participants who did not use rules of behavior compatible with centration, suggesting that consistency is partially driven by the development of a common skill.

We showed that the development of choice transitivity was not related to the development of transitive reasoning and it was only partially explained by the development of

attentional control. Instead, it was supported by the improvement of self-knowledge of preferences, and our design allowed us to identify different trajectories across domains. In the Social domain, children learned to select their most-preferred option. In the Risk domain, children learned to avoid their least-preferred option. In the Goods domain, children learned both. Overall, a main difference between adults and children was that adults consistently knew what they wanted but children did not. Said differently, children's imperfect self-knowledge of preferences was what made them act in an inconsistent fashion and differences across domains were due to asymmetries in preference awareness. The fact that transitive decision-making is promoted by self-knowledge of preferences and attentional control is compatible with recent developments in neuroscience that identifies value and attention systems from which economic decision making stems. The domain-specific biases revealed in the developmental trajectories and also in the adult performance (adult know best what they like in the Social domain and what they dislike in the Risk domain) suggest that values are acquired differentially across contexts.

The article is organized as follows. Section 2 presents the experimental design and the procedures we used. Section 3 gathers the theory and main hypotheses. Sections 4 and 5 report evidence that the developmental trajectories of consistent decision-making are domain specific. Section 6 addresses the contribution of developmental paradigms to these trajectories and presents some extensions. Section 7 concludes.

2 Experimental design and procedures

The experiment was conducted through tablet computers and the tasks were programmed with the Psychtoolbox software, an extension of Matlab.

2.1 Participants and goods.

We recruited 134 children from kindergarden (K) to 5th grade from Lycée International of Los Angeles, a bilingual private school. We ran 18 sessions, each with 5 to 10 subjects and lasting between 1 and 1.5 hours. Sessions were conducted in a classroom at the school. In all of our tasks, children made choices involving goods. Since taste significantly varies by age and gender and we wanted equally salient decisions, for the purpose of selecting the goods we divided our sample into four groups by age (K to 2nd and 3rd to 5th) and gender (male and female). We pre-selected 20 to 30 toys and stationary items. From that list, we screened seven highly desirable goods for each of the four groups, with some overlapping across groups. As a control, we ran 7 sessions with 51 undergraduate students (U). These were conducted in the Los Angeles Behavioral Economics Laboratory (LABEL) in the department of Economics at the University of Southern California. For the undergraduate

population, participants were recruited from the LABEL subject pool. Instead of toys or stationary, we made sure participants were hungry (they could not eat for two hours before the experiment) and we used snack foods. A summary description of the participants and goods is reported in Figure 1.[2]

	Male	Female	Total
K	12	7	19
1st	12	16	28
2nd	15	11	26
3rd	20	9	29
4th	11	8	19
5th	6	7	13
U	22	29	51

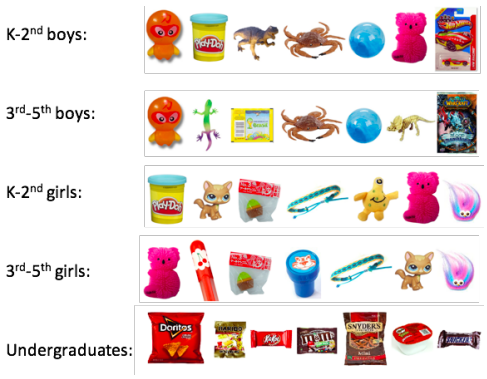

Figure 1: Description of participants (left) and goods (right) by age and gender.

2.2 Tasks.

Each participant completed a *Choice* task and a *Ranking* task with 7 "options" for each of the three domains *Goods*, *Social* and *Risk* (see section 2.3 for a detailed description of what the options are depending on the domain considered).

Choice task. In the Choice task, participants were presented with 21 trials which consisted of all 21 pairwise combinations of the 7 options. In each trial, one option was displayed on the left side and the other on the right side of the tablet's screen, as described in Figure 2 (left). A participant could select their preferred option by touching a button above the left or right options on their screen, or report indifference by touching a button between the two options. The 21 pairs of options were presented randomly, both in terms of trial order and the left-right presentation. The choice task was designed to measure the degree of transitivity of preferences. Indeed, assuming completeness of the preference ordering, the transitivity axiom states that if A is preferred over B and B over C, then A should be preferred over C. Finally, we added a 22$^{\text{nd}}$ trial, where subjects were presented with their most frequently and least frequently chosen options from the

[2] We believe it is crucial to have similar intensity in preferences across goods for all ages, even if it comes at the expense of having different goods.

preceding 21 trials.[3] This choice was tailored for each subject to be the simplest one (best against worst option) and allowed us to check for attentiveness and comprehension.[4]

Ranking task. In the Ranking task, participants received 7 cards, each with a picture of one of the options. Participants were instructed to rank these cards on a ranking board attached to their desk from most (right) to least (left) preferred. The ranking board had a green smiling face on the far right and a yellow neutral face on the far left, as described in Figure 2 (right).[5] Once a participant had finished ranking the options, an experimenter recorded her rankings on the tablet. While the Choice task was the main task to study transitivity, the Ranking task provided an *independent proxy* for ordinal ranking of options for each participant.

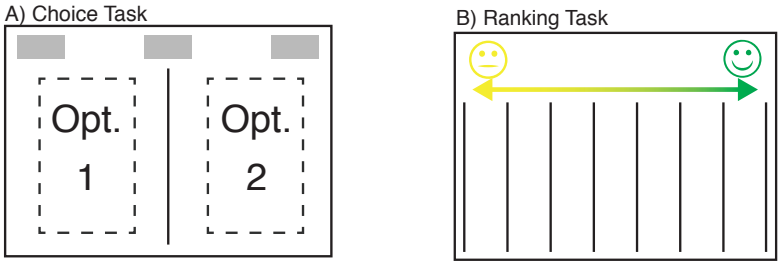

Figure 2: Visual presentation of the Choice (A) and Ranking (B) tasks.

Each participant also completed a *Transitive Reasoning* task that did not involve the seven options previously mentioned.

Transitive Reasoning task. The Transitive Reasoning task was designed to measure levels of transitive reasoning and to test the relationship between transitive choice and transitive reasoning abilities. Several tasks have been proposed in the literature (Bouwmeester and Sijtsma, 2006) but they typically require a significant memory or operational reasoning capacity. Since we wanted to avoid such elements, we designed a new visual test

[3] As the 21 pairwise choices were being made, the computer awarded 1 point to the option selected by the participant and 0.5 points to each option in case of indifference. After the 21 trials, the tally for each option was summed and the options with the most and least points were determined. In case of a tie, one of the options was chosen randomly.

[4] Procedures of this sort are common in psychology under the misleading terminology of "catch trials." They are an inexpensive method to check rather than assume that basic assumptions (such as attentiveness and understanding) are satisfied. They are infrequent in economics (see Charness, Levin, and Schmeidler (2014) and Brocas et al. (2019) for some exceptions).

[5] We explained to subjects that in the case that they liked two or more cards "exactly the same," they were to place the cards in the same area.

that does not require memory. The task consisted of seven questions of varying difficulty. Each of the seven questions consisted of two premises represented in two vignettes, and a third vignette with a response prompt. For each premise, participants were told that the animals shown in the vignette were at a party and the oldest wore a hat. Their task was to determine which animal in the third vignette was oldest and therefore should wear the hat. They could also select "?" to report they did not have enough information to know which animal was oldest. This is illustrated in Figure 3. Three of the seven questions did not require transitive reasoning and were included to test whether participants were paying attention. These are referred to as "pseudotransitivity" trials (Bouwmeester and Sijtsma, 2006). The remaining four did require transitive reasoning. Of the four transitivity questions, two of them were less difficult and two were more difficult.

Figure 3: Transitive Reasoning Task.

2.3 Domains: Goods, Social and Risk

A main objective is to study the ability of children to make transitive choices in different decision-making conditions. For this reason, we conducted the Choice and Ranking tasks in three domains: *Goods*, *Social* and *Risk*. The tasks were identical across domains (see Figure 2), except for the seven options involved. In the Goods domain, the options were the toys and stationary (or snacks for undergraduates) described in Figure 1 (right).

In the Social domain, the options were different sharing rules between tokens for oneself and tokens for another anonymous child of the same age and gender in a different school. We selected combinations that, according to previous studies of other-regarding preferences can be more or less valuable for some children depending on their predisposition for fairness, envy, spite, altruism and generosity (Fabes and Eisenberg, 1998; Fehr, Bernhard, and Rockenbach, 2008; Brocas, Carrillo, and Kodaverdian, 2017; Brocas and Carrillo, 2018a). Also, to make the decisions varied and interesting, we ensured that the "relative price of giving" changed from one option to the next. The seven options are depicted in Figure 4 (left), with tokens represented by filled circles. Each token mapped into either a toy (for children) or money (for undergraduates) as described in section 2.4 below.

A hand pointing inwards represented tokens for oneself and a hand pointing rightwards represented the tokens for the other person.

In the Risk domain, the options were binary lotteries. In each lottery, the participant could earn a number of tokens with a given probability and nothing otherwise. Tokens were again represented by filled circles which, just like before, mapped into either toys or money. Probabilities were represented by a two-color spinner wheel, with the green area corresponding to the probability of winning the token(s).[6] We selected six combinations with increasing quantities and decreasing probabilities, and different but reasonably close expected values (between 1 and 2.5). The riskiest and safest options had the lowest expected values, to put a cost premium on both extreme quantity (12 tokens) and extreme likelihood (probability 1). We added one dominated option (4 tokens with probability 1/2 should never be chosen over 5 tokens with probability 1/2) to test for task comprehension.[7] This option, however, should be preferred to most other options by risk neutral or moderately risk averse subjects. The seven options are depicted in Figure 4 (right). Notice that Goods, Social and Risk cover three of the most basic domains in the literature. They have been the object of a significant number of rationality studies in adults.[8]

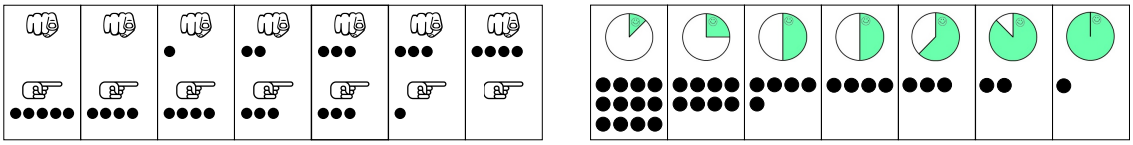

Figure 4: Options in the Social (left) and Risk (right) domains.

2.4 Timeline, incentives and implementation

All subjects completed the seven tasks in the same order: (1) Goods-Choice task, (2) Goods-Ranking task, (3) Transitive Reasoning task, (4) Social-Choice task, (5) Social-Ranking task, (6) Risk-Choice task and (7) Risk-Ranking task. All tasks were untimed. Upon completion of a given task by all subjects, instructions for the next task were given. We explained subjects that their rankings in the Goods-Ranking task were important to ensure that they would play with their preferred items in the rest of the experiment

[6] Previous experiments on risk with children have successfully employed a similar design (Reyna and Ellis, 1994; Harbaugh et al., 2002).

[7] Consistent with the idea that the majority of children understand tasks involving probabilities (Harbaugh et al., 2002), the dominated option is chosen between 4% and 11% of trials depending on the age group).

[8] Although we were also very interested in the Time domain (e.g., quantity x at date t), we did not include it for logistic reasons, as it was challenging to set up delayed payments in the school.

and would therefore collect the toys they liked best. We did not incentivize the other two Ranking tasks nor the Transitive Reasoning task.[9] By contrast, all three Choice tasks were implemented in an incentive-compatible way. From each Choice task, one trial (i.e., one pairwise comparison) was randomly selected by the computer and subjects received their selection in that trial. For the Goods-choice task, that meant one toy or stationary. For the Social-Choice task, we explained that each child had been paired with another student, of their same gender and grade level, from another school in Los Angeles (undergraduates were paired with a subject of another session, matched for gender), and that the selected sharing rule would actually be implemented.[10] For the Risk-Choice task, we explained that the spinner wheel in the selected trial would be spun by a blindfolded assistant. If the spinner arrow landed in the green part of the wheel, the subject would win the tokens associated to that choice. Otherwise, they would not win anything from that task. This was implemented by spinning a 12 inch spinner wheel for each child at a time at the end of each session in the front of the classroom.

To ensure the desirability of prizes, we announced that one token in the Social-Choice task corresponded to one unit of the good ranked first by the subject in the Goods-Ranking task, whereas one token in the Risk-Choice task corresponded to one unit of the good ranked second in the Goods-Ranking task. For undergraduate students, one token represented \$2.[11] Finally, at the end of the experiment we collected demographic information consisting of 'gender,' 'grade,' 'number of younger siblings,' and 'number of older siblings.' In addition to the variable payments, all participants also received a fixed show-up 'fee:' children received their highest ranked good in the Goods-Ranking task while Undergraduate students received \$5.

2.5 Methodology

Most recent experimental tests of rationality focus on the General Axiom of Revealed Preferences (GARP), where individuals choose between bundles of options given an implicit or explicit budget constraint, a system of prices, and a non-satiation assumption.[12]

[9] While incentivized mechanisms are definitely preferable, we decided against it in those tasks because, unlike for the Choice tasks, implementing them in an incentive-compatible way was challenging to execute and explain, increasing confusion among young children. In any case, all children took substantial deliberation time when completing the Ranking and Transitive Reasoning tasks. Also, it is important to notice that the Ranking tasks were used only as an independent measure of ordinal preferences and not to compute transitivity violations.

[10] The shared items were delivered to Foshay elementary school, a public school in the Los Angeles Unified School District.

[11] This imposed an extra constraint on toys, as we had to make sure that they were scalable.

[12] Prominent references include Harbaugh et al. (2001) (children population in the Goods domain), Andreoni and Miller (2002) (adult population in the Social domain) and Choi et al. (2007) (adult population

While the importance of these studies is obvious, we believe that, for the questions we are interested in, our design is methodologically superior. Here are the reasons. First, it is a more fundamental analysis of rational choice as it provides a direct (rather than an indirect) test of the transitivity axiom of preferences. Second, the extreme simplicity of the design (pairwise comparisons of options instead of system of prices and budget constraints) and the removal of assumptions such as non-satiation (sometimes violated in experimental studies) makes it especially suitable to test it in non-highly educated populations. This is especially important when we consider young children, who are susceptible to confusion, misunderstanding and inattention. Third, the presentation and design are virtually identical in the Goods, Social and Risk domains (see Figure 2 - left). Since the Choice tasks differ exclusively in the options considered, the number and severity of violations can be directly compared across domains.[13]

We collected data from all participants in the Goods-Choice and Goods-Ranking tasks. The tablets did not record the choices of two subjects in the Risk-Choice and Risk-Ranking tasks and one subject in the Social-Choice and Social-Ranking tasks.

3 Theory and hypotheses

3.1 Number of transitivity violations (TV)

A strength of the proposed design is its ability to provide a simple, interpretable and domain-free measure to determine the number of violations. For each triplet of options (A, B, C) a transitivity violation (TV) occurs if and only if there is a "cycle" in the three trials that directly compare the pairs of options: A is chosen over B, B over C, and C over A. The overall consistency of a subject is measured by (i) considering every combination of triplets of options (A, B, C) and (ii) for each triplet, determining whether there is a cycle in the three corresponding paired comparisons. With 7 options in our choice tasks, there are $7!/4!\,3! = 35$ triplets to consider.

It is key to realize that considering only triplets of options is enough to account for *all* TV. To illustrate this, suppose there is a a cycle involving four options: A chosen over B, B over C, C over D and D over A. This implies exactly 2 violations between triplets of options within those four. Indeed, if A is chosen over C, there is a violation

in the Risk domain).

[13] One caveat worth mentioning is that domains were implemented always in the same order (Goods followed by Social followed by Risk). By construction, Goods had to be performed first. Counterbalancing between Social and Risk was challenging for logistic considerations and difficult to implement given the small number of sessions per age group. Although we provided ample resting time between tasks and children were overall excited to play during the entire experiment, it is still possible that fatigue affected their performance.

within the triplet (A, C, D) whereas if C is chosen over A, there is a violation within the triplet (A, B, C). Analogously, if B is chosen over D, there is a violation within the triplet (A, B, D) whereas if D is chosen over B, there is a violation within the triplet (B, C, D). A similar argument applies for violations in cycles of five or more options. The only remaining issue is how to count violations when subjects express indifference. We decided to use a somewhat ad-hoc method and allocate 0.5 violation to triplets involving "weak" violations, that is, when we observe A chosen over B, B over C, and A indifferent to C.

Notice that a choice in a trial which results in a TV for a certain triplet of options, may preclude a TV for another triplet. For example, in the four-options case presented above, A chosen over C implies a TV for the triplet (A, C, D) but it also precludes a TV for the triplet (A, B, C). As a consequence, not all violations are equally severe. In the next subsection, we determine a way to measure severity of TV.

Finally, it is essential to put into perspective the number of violations empirically observed. As a benchmark of comparison, we determined through simulations the expected number of TV that subjects who played randomly would incur. These depend on the number of indifferences. Since our subjects express indifference between 5.1% and 17.3% of the trials depending on the age and domain (see Appendix A2 for details), we simulated random choices under indifference probabilities between 0.0 and 0.5 and found that, on average, subjects would incur between 9 and 11 violations.[14]

3.2 Severity of transitivity violations

Number of TV is only one possible way to address consistency. A complementary way is to look at severity. We compute two measures of severity of TV (also domain-free) for each subject. One measure, called *choice reversals*, counts the minimum number of choices that need to be reversed to restore transitivity in all choices. To compute that number, we designed an algorithm that sequentially changed the choices made in each trial, and computed the total number of TV after each change. If there were no TV at any point, the algorithm stopped. After all single trials had been exhausted, the algorithm repeated the procedure over pairs of trials, then triplets and so on. Thus, a score of n means that out of the 21 trials, a minimum of n choices needs to be changed in order to restore full consistency. The other measure, called *choice removals*, is similar except that it counts the minimum number of choices that need to be removed to restore consistency. We determined this value using an algorithm similar to that for counting choice reversals.

[14]For indifference probabilities above 0.5, the number of violations fall (trivially, in the limit where the subject expresses indifference between all pairs, there are no TV). However, none of our subjects expressed indifferences in that range.

3.3 Hypotheses

Our goal is to study consistency across age groups and domains. We test four hypotheses.

The first hypothesis concerns consistency across age groups. It is based on previous research on GARP consistency in children and adults.

Hypothesis 1 *In all domains, the number of transitivity violations is high for the youngest children, monotonically decreases with age and is close to zero in the adult population.*

The second hypothesis is related to consistency across domains. Comparing violations in Goods, Social and Risk is possible thanks to our design, where every element of the experimental procedure and the statistical analysis is identical across domains, except for the options presented to the subjects. We conjecture that if the development of consistent decision making is exclusively related to the evolution in the cognitive capacities of children (attention, centration, reasoning and logical thinking), it should depend on the intensity of preferences over the alternatives but not on the nature of the options involved. We should therefore observe the same pattern independently of the domain.

Hypothesis 2 *For each age group, the number of transitivity violations is similar across domains.*

Our third hypothesis is about consistency as a function of the item's rank as explicitly revealed in the Ranking task. It is based on the intuition that it is easy for participants to know what to choose in all domains when the options they like most or least are present, and when the options involved are ranked very differently. By contrast, decisions in all domains are difficult when choices involve options that are close to each other and they do not involve the options they like the most or the least. This hypothesis is compatible with current research in cognitive psychology and neuroscience (Bogacz, Wagenmakers, Forstmann, and Nieuwenhuis, 2010; Gold and Shadlen, 2001), which argues that comparisons between close and intermediate options are relatively more difficult to make, resulting in more time to deliberate and more choice reversals.

Hypothesis 3 *In all domains and age groups, transitivity violations are more frequent for options with intermediate ranking and options that are ranked close to each other than for options with high or low ranking and options that are ranked far apart.*

Our final hypothesis traces consistency to the existing developmental paradigms. We focus on the three paradigms most likely to impact decisions in our experiment given our age groups: (i) attentional control, an ability that allows to focus on choices to best

evaluate them; (ii) centration, the tendency of young children to not evaluate and combine multiple attributes of a choice simultaneously; (iii) transitive reasoning, a logical ability that may sustain transitive choice. Children in different age groups are known to exhibit different patterns of behavior with respect to these three paradigms. Attentional control develops over childhood but remains significantly less developed in children in our window of observation than in adults (Davidson et al., 2006; Astle and Scerif, 2008). Centration is often demonstrated in children between the ages of 2 and 7, a developmental stage called preoperational (Crain, 2015). Transitive reasoning is acquired gradually and as a function of the difficulty of the inference to draw (Bouwmeester, Vermunt, and Sijtsma, 2007).

Hypothesis 4 *The development of choice consistency across domains is explained by changes in attention, centration and logical reasoning.*

The analysis in the next sections tests these four hypotheses. The results summarize our findings.

4 Transitivity violations by age and domain

Our first cut at the data consisted in studying violations by grade and domain. Formally and following the theory described in section 3.1, we first counted the number of times choices between triplets of items were intransitive for each subject in each Choice task. Then, we computed the average number of violations by grade and Choice task. The results are summarized in Figure 5.[15]

In the Goods domain, there was a significant improvement up to **4th** grade. Children in **K** had significantly more violations than children in all other grades (two-sided t-test, all p-values < 0.005), children in **1st grade** had significantly more violations than children in **4th** grade and above ($p < 0.01$) and children in **2nd** and **3rd** grades had significantly more violations than subjects in **U** ($p < 0.01$). However violations in **4th** grade, **5th** grade and **U** were not significantly different from each other ($p > 0.4$). In the Social domain, we observed a similar pattern. TV were significantly higher in **K** compared to all grades above **3rd** ($p < 0.02$). Children in **1st** and **2nd** grade committed more violations than adults ($p < 0.03$). There was no statistically significant differences between TV scores in **3rd** and higher grades. In the Risk domain however, there was no improvement, with no group of children exhibiting a number violations significantly different from each other or

[15] As discussed in section 3.1, random players would commit between 9 and 11 violations, above the actual numbers obtained even among the most inconsistent of our participants. This result is in line with earlier literature on consistency in children (Harbaugh et al., 2001).

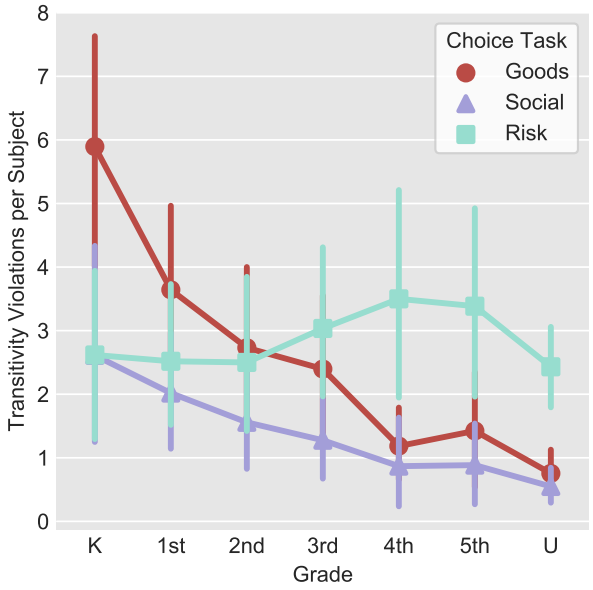

Figure 5: Average number of TV in the Choice tasks (y-axis) for each grade (x-axis) and each domain (Goods - red circle; Social - purple triangle; Risk - green square).

different from adults.[16]

The comparison of TV across domains delivered the following conclusions. Although all groups had more violations in the Goods than in the Social domain, differences were significant only for **K** ($p < 0.004$) and **1st** ($p = 0.033$). The comparison between Risk and Goods is more intriguing: the youngest children had significantly fewer violations in the former ($p = 0.02$) while the adults exhibited the opposite pattern ($p < 0.001$). This is the result of a significant improvement with age in the Goods domain together with no change in the Risk domain. Finally, there was no significant differences between the number of violations in the Risk and Social domains among children up to **2nd** grade. All older participants exhibited significantly more violations in the Risk than in the Social domains ($p < 0.003$). To sum up, TV monotonically decreased with age in the Goods and Social domains (although at different rates), but not in the Risk domain. Participants in **U** were not making significantly more transitive choices compared to participants in **4th** or **5th** grade in any domain, initially suggesting that the development of transitive decision-making is domain-dependent but stops around 10 years of age in all cases. Also, the behavior of our adult population was different across domains, featuring almost no

[16] In particular, violation counts were not significantly different between **4th** and **U** (p-value = 0.19) or between **5th** and **U** (p-value = 0.31).

violations in the *Goods* and *Social* domains but a positive and statistically significant number of violations in the *Risk* domain. This indicates that the differences observed in the children population persist over time.

We then studied the severity of violations. Following the procedure described in section 3.2, we determined for each age group and domain the number of choices that needed to be reversed to restore transitivity.[17] The results are depicted in Figure 6.

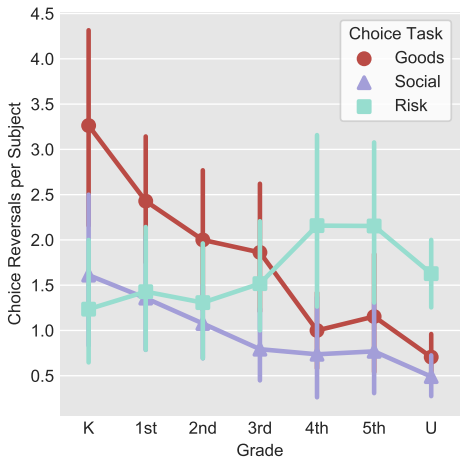

Figure 6: Severity of violations across domains and age: minimum number of choices that need to be reversed to restore transitivity.

The evolution over age groups was remarkably similar to the evolution in the number of TV, and all three measures were highly correlated with each other in all three Choice tasks (Pearson > 0.90, Spearman > 0.90, $p < 0.001$).

The results of this section are summarized as follows.

Result 1 *We found support for Hypothesis 1 in the Goods and Social domains but not in the Risk domain: consistency (amount and severity) improves with age in Goods and Social but not in Risk. In all domains, violations are similar for 4th-5th graders and college students (Figures 5 and 6).*

Result 2 *We did not find support for Hypothesis 2. Consistency is domain-dependent. It is weakly higher in Social than in Goods for all age groups. It is higher in Risk than in Goods for younger children and lower for older children (Figures 5 and 6).*

The results also indicate that while there are significant differences between our youngest

[17] We also considered the number of choices that needed to be removed to restore transitivity. The results were very similar.

(5-7 years old) and oldest (9-11 years old) children, changes are gradual with age and do not exhibit jumps (with rare exceptions). We also observe few differences between **K** and **1st** (except in the Goods domain), between **2nd** and **3rd** or between **4th** and **5th**. This is consistent with known developmental stages and sub-stages, as emphasized in the developmental psychology literature (Harris and Butterworth, 2012). The period ranging from 5 to 6 years old (**K** and **1st**) corresponds to the end of the Piagetian pre-operational stage, where children have not yet developed concrete logic. The period ranging from 7 to 10 years old corresponds to the concrete operational stage, which prolongs itself until adolescence and features the gradual acquisition of logical thinking. The development of executive functions, which sustain working memory abilities, is marked by an acceleration around 8 years old: children in **2nd** and **3rd** grades perform significantly worse at working memory tasks than children in **4th** and **5th** grades (Brocki and Bohlin, 2004). The differences and similarities across grades also reflect the french education system of this school, where students are separated in education cycles: "cours élémentaire" CP-1 and CP-2 (basic course 1 and 2, which correspond to 2nd and 3rd grade) and "cours moyen" CM-1 and CM-2 (middle course 1 and 2, which correspond to 4th and 5th grade). For the remaining of the paper, we group children in 3 age-categories **K-1st** (47 subjects), **2nd-3rd** (55 subjects) and **4th-5th** (32 subjects). This facilitates comparisons as it limits the number of groups to consider and increases the statistical power of the tests.[18]

5 Transitivity violations within domains

5.1 Transitivity in the Goods domain: the developmental template

To investigate systematic patterns of transitivity violations in the Goods domain, we employed the following methodology. For each individual, we used the Goods-Ranking task to rank the seven options (from most (1) to least (7) preferred). We then used the Goods-Choice task to determine a violation score as follows. For each triplet of options (A, B, C), we assigned a value of 1 to all 3 pairs of options (A-B, A-C, B-C) if they were involved in a TV and a value of 0 to all three pairs if they were not. We then determined the violation score of each pair as the percentage of times the pair was involved in a triplet that exhibits a TV. Each pair of options was then characterized by its violation score. Finally, we took the options ranked a and b for all the individuals in an age group

[18] Notice that some subjects from the same age-group were playing for different toys (males and females, 2nd and 3rd graders). Our main concern was to find seven high-value options for each age and gender so that comparisons are meaningful.

and we determined the average violation score in that age group for that pair.[19] Figure 7 represents the color-coded result of this exercise, with the higher ranked option in each pair presented in the x-axis and the lower ranked option in the y-axis. Darker colors reflect fewer violations.

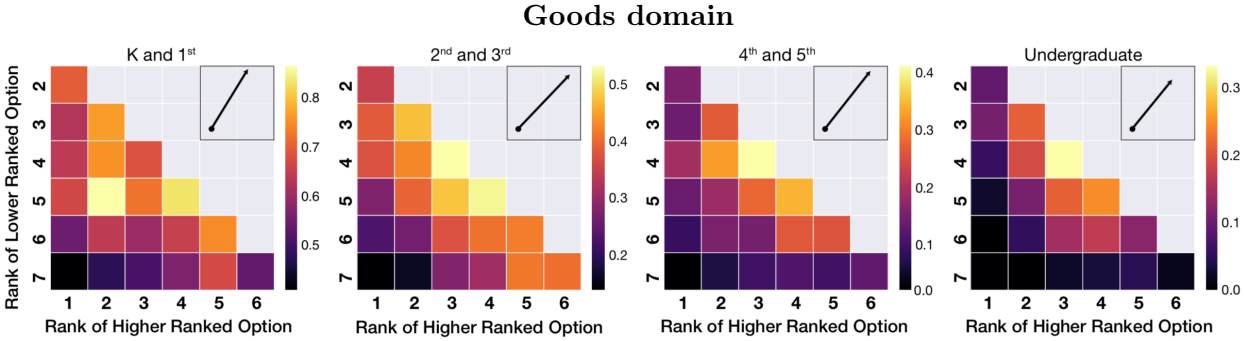

Figure 7: Heatmap of TV in the Goods domain as a function of ranking. The color of cell (a, b) represents the average fraction of TV involving the goods ranked in a^{th} and b^{th} position in the Goods-Ranking task. The vector in the upper-right corner captures the gradient of the increase in TV as the rankings of the two goods become closer to each other (through a decrease by one rank of the higher ranked option a (x-axis) or an increase by one rank of the lower ranked option b (y-axis)).

The heatmap can also be used to study *marginal sensitivity to violations*. One can expect that when choices are easy to make, inconsistencies are less prevalent. Easiness might be a matter of picking what we like, or it might be a matter of avoiding what we do not like. If consistency is only driven by choosing what is liked, it should not matter what the lower-ranked option is, and if consistency is only driven by avoiding what is liked least, then it should not matter what the higher-ranked option is. We tested which choices were harder to make by analyzing how consistency changed if we marginally changed the rank of the higher- or lower-ranked option. From Figure 7, if participants exhibited a certain level of consistency with the pair of options ranked (a, b), we could study the change in consistency by moving one cell to the right, that is, to $(a + 1, b)$ thereby decreasing the higher-ranked item without changing the lower-ranked item. Analogously, we could also study the change in consistency by moving one cell up, that is, to $(a, b - 1)$, thereby increasing the lower-ranked item without changing the higher-ranked item.

Formally, we considered every pair of adjacent boxes in the same row and we determined the difference in the violation score between a box and the box to the right of it. We then computed the average difference over all pairs of boxes and reported this number

[19] It is crucial to notice that rank a corresponds to different items for different individuals since each participant has his own personal ranking.

as the left-right gradient of the vector in the top-right corner of Figure 7 for each age group. A vector with a larger x-coordinate means a greater increase in violations when the higher-ranked option is decreased by one rank. We did the same with every pair of adjacent boxes in the same column to determine the up-down gradient. In this case, a vector with a larger y-coordinate means a greater increase in violations when the lower-ranked option is increased by one rank.

Finally, we use the data from all 15 vectors in all 4 age-categories to estimate a bivariate kernel density. Figure 8 depicts a plot in polar coordinates of this data.

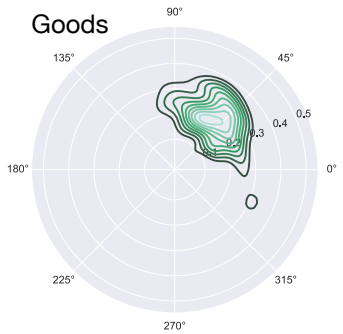

Figure 8: Sensitivity of violations to ranking in the Goods domain (contours are steps of equal height of the kernel density estimate of the gradient vectors in all age-categories).

Although consistency in the Goods-Choice task improved significantly with age, that improvement was not uniform across all paired choices. Trials featuring options ranked very differently were unlikely to be involved in a TV (darker cells). By contrast, trials featuring options ranked similarly were significantly more likely to be involved in TV (lighter cells). This general pattern was independent of age. It can best be grasped by looking at the average gradient of the heatmap (vectors in the top-right corners of Figure 7) as well as the polar plot representation in Figure 8: the gradients close to the 45° line indicate more violations if the rank of the higher-ranked option decreases and if the rank of the lower-ranked option increases. A t-test confirmed that both the x- and y-coordinates were positive and significantly different from 0 (p-value < 0.01) in all age groups, with the exception of the x-coordinate in **K-1st**.

We also observed convergence to a state where 4th-5th grades and undergraduates almost never committed TV when choices involved their best (left column) or their worst (bottom row) options. Instead, the majority of their TV were concentrated in items with intermediate ranks (3rd, 4th and 5th).

Overall, the behavior observed in this domain was consistent with the hypothesis that participants made choices by estimating and comparing noisy values. Under that

hypothesis, decisions involving options close in value are confusing and prone to error, while decisions involving options valued differently are easy to make. The developmental trajectory also suggests that the evaluation process becomes less and less noisy over time, reducing the number of confusing decisions, and hence the number of violations, especially among options ranked very differently. We summarize the results as follows.

Result 3a *We found support for Hypothesis 3 in the Goods domain. In all age groups, transitivity violations are more prevalent when the options involved have intermediate and similar ranks: children learn over time to know both what they like most and what they like least (Figures 7 and 8).*

5.2 Transitivity in the Social domain: the effect of centration

We did not expect to find that small children would be significantly more consistent in the Social-Choice task compared to the Goods-Choice task (Figure 5). A possible explanation is that goods are atomic and need to be evaluated as a whole. So, even though it is possible to focus on one characteristic of the good (size, shape, color, etc.), it is more probable that our subjects look at toys as units that are compared to each other simultaneously on all dimensions. By contrast, social options may be naturally decomposed into several simple attributes such as "objects for self," "objects for other," and "total number of objects." Then, while a rational decision-maker should trade-off the relative importance of each attribute, it is plausible that some participants focus on one single attribute at a time (centration) and use simple lexicographic strategies to make decisions. Such behavior is likely to be more prevalent among our youngest population, as it is well-known that children below 7 years of age have difficulties in reasoning simultaneously over multiple attributes (Piaget, 1952; Crain, 2015). Interestingly, this process would "buy" consistency, as it is trivial to follow and unlikely to produce transitivity violations.

To explore this hypothesis, we retained three attributes –reward to self, reward to other, and total reward– and three rules –pick the maximum, pick the minimum, or be indifferent. We considered every possible strategy where the participant applied a rule to a single attribute and, if two or more options were equivalent under that rule, moved to a second attribute and applied the same or a different rule. Despite the large set of possible strategies of this sort, only three were ever used by our subjects: "maximize amount for self then minimize amount for other," "maximize amount for self then maximize amount for other" and "maximize amount for self irrespective of amount for other." We counted all participants who complied with one of these three strategies and those who made exactly one mistake. We called these participants 'simple policy users.' We noticed that the first policy (the most popular of all three) was consistent with the choices of 35% of **K-1st**,

20% of **2nd-3rd**, 22% of **4th-5th** and 4% of **U**. We then removed all simple policy users from our sample, leaving us 125 participants, and compute TV by age group after the exclusion. Figure 9 presents the result of this exercise.

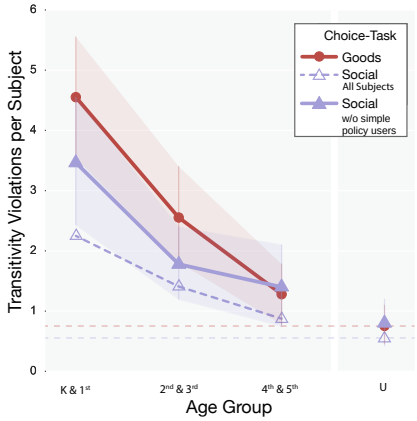

Figure 9: Comparison of TV in Goods and Social domains with (dashed line) and without (full line) simple policy users.

When we removed simple policy users, the developmental signature matched that of the Goods domain. Participants in the **K-1st** age group were significantly more inconsistent than all other participants (all p-values < 0.01). Participants in the **2nd-3rd** age group had significantly more violations than participants in the **U** age group (p-value $= 0.014$) and participants in the **4th-5th** age group were not significantly different from participants in the **U** age group (p-value $= 0.121$).

It is important to emphasize that while the behavior of simple policy users may come from shortcut heuristic reasoning, it may also reflect the underlying preferences of fully cognitive subjects. Our data cannot distinguish between these two radically different but observationally equivalent decision processes. However, the higher proportion of those choices by children in the youngest age group is suggestive of the former process, at least for some subjects: participants who have not yet overcome centration are more likely to pick a simple policy that makes them "look consistent" to the outside observer.

Next, we repeated in the Social domain the heatmap analysis performed on the Goods domain using only the participants who were not simple policy users. The result is reported in Figure 10.[20]

A comparison between Figures 7 and 10 revealed systematic differences between the

[20]Notice that the heatmap with all participants is qualitatively very similar (roughly a re-scaling of Figure 10 with all areas darker) since simple policy users have no or almost no transitivity violations for any triplet of options.

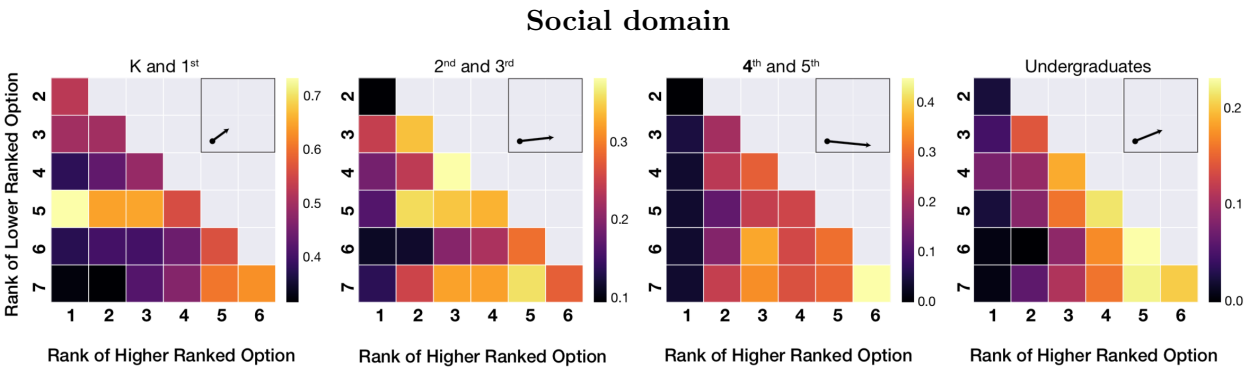

Figure 10: Heatmap of TV in the Social domain as a function of ranking and excluding simple policy users.

Goods and Social domains in the evolution of transitivity violations between similarly- and differently-ranked options. More precisely, in the Social domain, children learned to become highly consistent in choices involving their best options ("they learn to know what they like most") but they committed a substantial number of violations in choices involving their worst options, even when compared to medium or high ranked ones ("they do not learn to know what they like least"). This was true even in the **U** age group.

We then performed the same marginal sensitivity analysis as in the Goods domain (again after removing simple policy users). Reinforcing the developmental trajectory highlighted above, we noticed that the x-coordinates were positive and significantly different from 0 (p-value < 0.01), overall and for all age groups except for participants in the **K-1st** age group. However, and contrary to the Goods domain, the y-coordinates were not significantly different from 0 for any age group. This confirms that our participants had a very clear idea of what they liked best but a much less clear idea of what they liked least. The result can also be seen in the polar plot representation of the vectors provided in Figure 11.

The results of this section can be summarized as follows.

Result 3b *We did not find support for Hypothesis 3 in the Social domain. In all age groups, violations are more prevalent for options with low rank than for options with high rank: children learn over time to know what they like most but not what they like least (Figures 10 and 11). Also, consistency in the Social domain is aided by simple rules.*

5.3 Transitivity in the Risk domain: too complex

As in the Social domain, making choices in the Risk domain requires the evaluation of multiple attributes. We hypothesize that centration can also play a role in this domain.

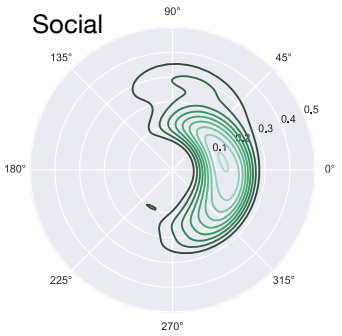

Figure 11: Sensitivity of violations to ranking (Social domain)

We considered two attributes per option –reward amount and probability–, and three simple rules –pick the maximum, pick the minimum, or be indifferent. Although this results in a large number of simple policies, only four were effectively used, and one was overwhelmingly predominant. Indeed, among the 56 subjects who used simple policies, 89% chose to "maximize the amount irrespective of the probability." This corresponded to 44% of **K-1st**, 35% of **2nd-3rd**, 19% of **4th-5th** and 8% of **U**. It means that, just like in the Social domain, simple policies were more prevalent in the younger population. It also supports the previously mentioned idea that centration results in heuristic rules that are conducive of transitivity.[21]

We removed from our sample all participants whose behavior was consistent with a simple policy, leaving us with 128 subjects, and analyzed transitivity violations in this subsample. The results are reported in Figure 12.

The developmental signature *did not* match that of the Goods and Social domains. At the same time, it is less puzzling than when we considered the full sample. Indeed, violations by the **U** group were significantly smaller than those by any younger group: **K-1st** (p-value = 0.002), **2nd-3rd** (p-value = 0.005) and **4th-5th** (p-value = 0.015). However, all school-age groups performed at similar levels from each other. The result indicates that a potential milestone for Risk is outside our window of observation, somewhere in middle or high school. It also suggests that trading-off amounts and probabilities is cognitively harder than trading-off gains for oneself and gains for others or evaluating attributes of a good.

We also repeated the heatmap (Figure 13) and marginal sensitivity analyses (Figure

[21] As before, it is not possible to distinguish between simple policy users and rational (in this case, risk-loving) individuals. Once again, the difference across age groups in the number of subjects following such strategy is an indication that centration and heuristic decision making is a reasonable explanation for at least some of those subjects.

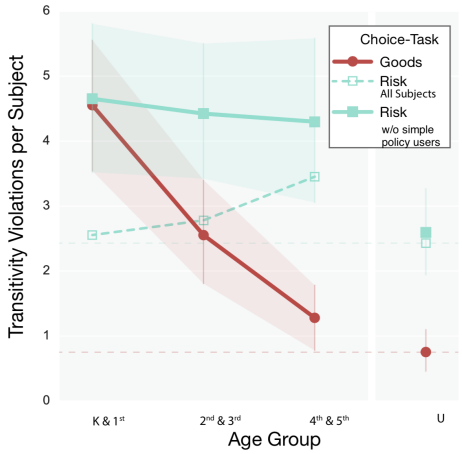

Figure 12: Comparison of TV in Goods and Risk domains with (dashed line) and without (full line) simple policy users.

14) after removing simple policy users.

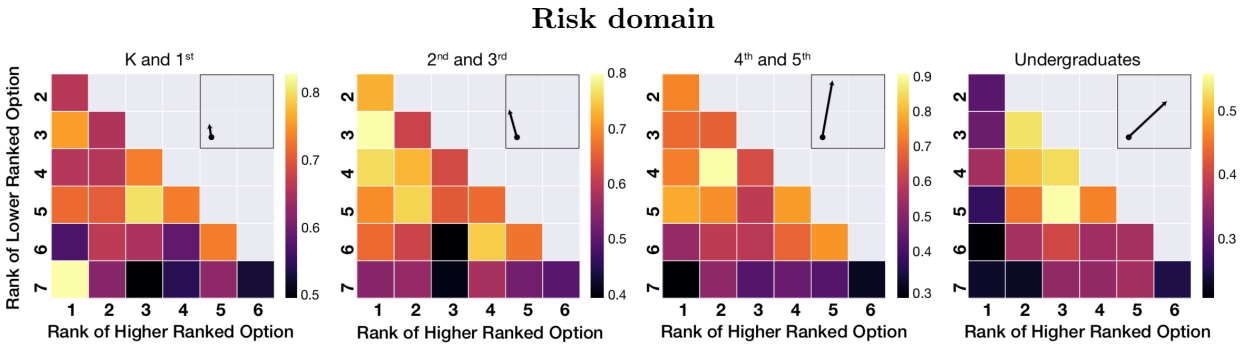

Figure 13: Heatmap of TV in the Risk domain as a function of ranking and excluding simple policy users.

As we can see from the heatmap, the results were the opposite to those we obtained in the Social domain: children learned to become significantly more consistent in choices involving their worst options but less so in choices involving their best options ("they learn to know what they like least but not what they like most"). Similar to the Social domain, however, older children were better able to make integrative decisions, in this case, trade-offs between reward amounts and probabilities.[22] The conclusions were confirmed with the sensitivity analysis, where the y-coordinates were significantly different from 0 overall

[22]This is made more evident in the analysis of the evolution of preferences reported in Appendix A3.

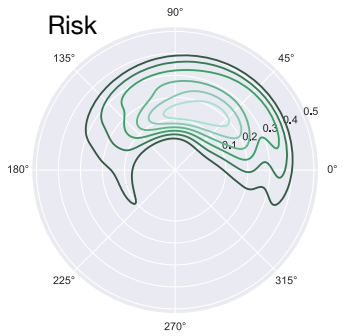

Figure 14: Sensitivity of violations to ranking (Risk domain)

and for two age groups (**4th-5th** and **U**, t-test p-values < 0.05), and the x-coordinates were not different from 0 in any age group. The result is summarized as follows.

Result 3c *We did not find support for Hypothesis 3 in the Risk domain. In all age groups, transitivity violations are more prevalent for options with high rank than for options with low rank: children learn over time to know what they like least but not what they like most (Figures 13 and 14). Consistency of children in the Risk domain is also aided by simple rules. Among those who do not use simple policies, performance does not improve over our age window, but it is worse than in the college population.*

5.4 Are transitivity violations correlated across domains?

While the previous sections show that the developmental trajectory of consistency is different across domains, it does not address the issue of whether participants who commit more violations in one domain are also likely to commit more violations in a different domain. Said differently, we would like to know whether consistency is driven by a common factor or if it results from the development of domain-specific skills.

When we considered the full sample, we found that TV in the Goods and Risk domains were not correlated with one another, while TV in Goods and Social or TV in Risk and Social were (Pearson coefficient 0.38 and 0.30, respectively; p-value < 0.0001 for both). When we removed simple policy users, TV were highly and very significantly correlated across all domains (Pearson coefficient $= 0.36$, p-value $= 0.0006$ between Goods and Risk, Pearson coefficient $= 0.43$, p-value < 0.0001 between Goods and Social, and Pearson coefficient $= 0.49$, p-value < 0.0001 between Risk and Social). This strong correlation suggests that participants' consistency is partially driven by the development of a skill useful in all domains.

5.5 Summary and discussion

The results in sections 4 and 5 reveal that consistent choice improves with age only in two domains, Goods and Social, and that developmental trajectories are different across all three domains. We have also shown that only the Goods domain follows the expected developmental template, featuring inconsistencies more prevalent among options with intermediate ranking and options that are ranked close to each other. However, we have also shown that inconsistencies are more prevalent for lower ranked options in the Social domain whereas they are more prevalent for higher ranked options in the Risk domain. These differences are intriguing. The fact that choices involving options ranked as extreme may be difficult to make in a consistent manner suggests that participants have imperfect knowledge of their preferences. Finally, individuals who exhibit inconsistencies in one domain are also more likely to exhibit inconsistencies in the others. This indicates that a common skill may be responsible for the improvement of consistent decision-making in all domains. Interestingly, current research in neuroscience argues that the attention system is recruited when subjects make comparisons between options (Krajbich, Lu, Camerer, and Rangel, 2012). Therefore, a candidate for this common skill is the developing ability to dedicate attention to choices.

6 The determinants of transitive choices

The objective of this section is to investigate in more detail the possible underlying mechanisms that promote consistency. To this end, we test Hypothesis 4 by studying the relationship between consistent choices and known developmental changes in centration, attention, and transitive reasoning. Given the results of section 5, an important question is whether the development of self-knowledge of preferences is a paradigm in itself that contributes independently to consistency. We address that question by providing a measure of self-knowledge of preferences and by analyzing its relationship with the other variables of interest.

6.1 Centration

In our experimental paradigm, centration has been identified as a simple lexicographic policy followed by the participant in every pairwise comparison. It is therefore immediate that centration is, by construction, strongly associated with consistency.

However, from the 59 subjects who used simple policies in the Social domain and the 55 who used simple policies in the Risk domain, only 19 used simple policies in both domains. We also compared TV in the Goods domain (where no simple policies are available) by

subjects who did and did not use simple policies in the other domains and found no significant differences. This means that centration is not a systematic characteristic for some children. Also, those who exhibit centration in some domain are not necessarily more inconsistent when such simple strategies are not available.

6.2 Self-knowledge of preferences

From a participant's set of decisions in a Choice task, it is possible to extract his implicit (or revealed) ranking for the options in that task. We computed this revealed ranking in a very simple way, by tallying the selections "for" and "against" each of the options as briefly explained in footnote 3: each time the participant selects an option, 1 point is added to the running tally of that option and when the participant expresses indifference, 0.5 points are added. The tallied points for each option are summed, and the options are ordered according to this sum, giving the subject's implicit ranking.[23]

This implicit ranking can then be compared to the explicit ranking determined in the Ranking tasks. For each participant and task, we computed a measure of distance between the implicit and explicit rankings again using a simple procedure. We assigned to each option a score representing the absolute difference between its rank in the implicit and explicit rankings (e.g., if A is ranked 3rd according to the explicit ranking and 2nd according to the implicit ranking, its score is $|3-2|=1$). We then computed the average score of all options as each subject's discrepancy score.[24] Figure 15 depicts the average discrepancy score by domain and age group. As can be seen from the graph, in the Goods domain, the ability to choose according to one's explicitly disclosed preferences developed gradually. The same was true in the Social domain for age groups **2nd-3rd** and above. By contrast, discrepancies remained constant over time in the Risk domain, suggesting a persistent inability to draw choices from explicit preferences. Overall, the result reinforces the idea that preferences are not well established in young children. This has consequences both in terms of transitivity violations as well as in discrepancies between choices and rankings. This disparity decreases with age in the Goods and Social domains but not in

[23] Another method to elicit such ranking is to set up a random utility model (RUM) and to estimate the value of each option for each child. Our study was however not designed to extract the necessary number of observations to make that procedure successful. This would have required to increase substantially the length of the session, and would have likely come at the expense of data quality. Moreover, the amount of noise in the RUM model is usually normalized. In our case, it is likely that the amount of noise around choice differs across populations and it should be estimated (by allowing the scale of the extreme value distribution to vary across age groups). Unfortunately, values and noise parameters cannot be jointly identified, limiting further the value of the exercise.

[24] This ad-hoc method implicitly imposes cardinality, as it treats equally the distance between any two consecutive ranks. We view it as a simple, non-axiomatic way to capture the closeness of the two methods in ranking items.

the Risk domain.

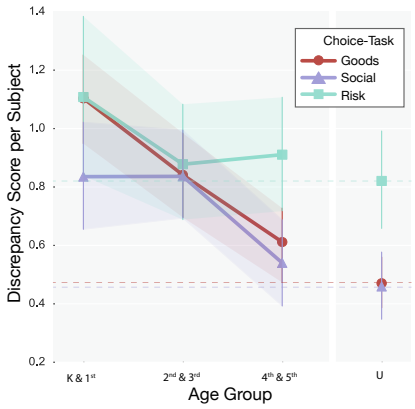

Figure 15: Discrepancies between explicit and implicit rankings (all subjects).

Interestingly, a high number of TV was associated with a high discrepancy score between implicit and explicit ranking in the Goods (Pearson = 0.61), Social (Pearson = 0.54) and Risk (Pearson = 0.60) domains, all p-values < 0.0001. In other words, while participants did not seem to have trouble choosing between pairs of alternatives (Choice tasks) or reporting a ranking among all options (Ranking tasks), the preference elicitation method had a significant impact on the reports of those committing many TV. Differences between implicit and explicit rankings indicated that participants did not have a unique, unambiguous way of ranking options. Said differently, they did not know their preferences well enough to produce similar classifications. The fact that differences between rankings were associated with TV suggests that imperfect knowledge of preferences was a likely driver of violations in the Choice tasks.[25]

6.3 Attention

Remember that after all 21 paired comparisons in a Choice task we included an "attention trial", more precisely, a last comparison between the subject's most-favorite and least-favorite options, as revealed by their choices. This decision provides a measure of task comprehension and focus. By this measure, most children in our sample were attentive. Indeed, 70% in **K-1st**, 76% in **2nd-3rd**, 84% in **4th-5th** and 100% **U** chose their (implicitly ranked) first option over their (implicitly ranked) last option in all three

[25]This interpretation is consistent with a random utility theory in which noise affects the perception of true underlying values. A decision-maker who does not know his preferences well is subject to larger perceptual mistakes. This causes transitivity violations in Choice tasks and different answers to similar questions producing different rankings across contexts.

domains. Most of the participants who did not, expressed indifference in one domain and answered "correctly" the other two, and no children answered incorrectly all three. Interestingly, TV and performance on attention trials was correlated in all three Choice tasks: Goods (Pearson = 0.54, Spearman = 0.43), Social (Pearson = 0.40, Spearman = 0.20), and Risk (Pearson = 0.36, Spearman = 0.32), all p-values < 0.01. These results suggest a relationship between ability to choose consistently and attention mechanisms.

In the same lines, the discrepancies between explicit and implicit rankings discussed above are correlated in all domains with performance in attention trials: Goods (Pearson = 0.38, Spearman = 0.34) Social (Pearson = 0.32, Spearman = 0.21) and Risk (Pearson = 0.27, Spearman = 0.30), all p-values < 0.01.

Taken together, the results highlight the importance of focus or attentiveness in achieving transitivity. While it is not overly surprising, especially in such a young population, it is a result worth noting.

6.4 Transitive reasoning

Remember that in the Transitive Reasoning task we proposed a novel, seven-question design with three levels of difficulty to assess transitive reasoning abilities. We hypothesize that transitive reasoning ("if A is older than B and B is older than C, who is older A or C?") could be correlated with transitive choices ("you expressed a preference of A over B and B over C; do you prefer A or C?"). Indeed, while these questions are conceptually very different, they still share some common aspects in the reasoning process.

We counted for each participant the number of mistakes accumulated in each level of difficulty of the transitive reasoning task. As expected, in all three levels, the **K-1st**, **2nd-3rd**, and **4th-5th** groups accrued significantly more errors than the **U** group (p-value < 0.001, p-value < 0.05, and p-value < 0.05, respectively). Participants in the **K-1st** group made more mistakes on the most difficult reasoning trials than they did on the easy or medium trials (p-value = 0.02 and p-value = 0.0001, respectively). Within the other age groups the average error, counts were similar across trial difficulty (with the exception of the **2nd-3rd** group which had more mistakes on easy than on difficult trials (p-value = 0.02)).

We also noticed that performance in the transitive reasoning task was correlated with attentiveness. This was true for the most difficult trials (Pearson = 0.20, p-value < 0.01) as well as for all levels of difficulty together (Pearson = 0.22, p-value < 0.01). Finally, performance in the transitive reasoning task was correlated with the level of discrepancy between implicit and explicit rankings in all domains: Goods (Pearson = 0.33, p-value < 0.0001), Social (Pearson = 0.20, p-value < 0.01) and Risk (Pearson = 0.17, p-value < 0.05). Overall, transitive reasoning was associated with the same explanatory variables

as transitive decision-making.

6.5 Relationship between TV and developmental variables

Several factors can potentially be correlated with TV. Table 1 presents descriptive statistics by age group of the variables that, according to our previous analysis, are likely to affect choice transitivity. These include the number of mistakes in the transitive reasoning task (*TR mistake*), the number of mistakes in the attention trials (*Attention*) and the discrepancy score between the decisions in the Choice and Ranking tasks in each domain (*Discrepancy*).

	TR mistake	*Attention*	*Discrepancy*		
			Goods	Social	Risk
K-1st	2.06 (0.21)	0.21 (0.05)	1.10 (0.08)	0.84 (0.09)	1.11 (0.14)
2nd-3rd	1.22 (0.14)	0.19 (0.05)	0.84 (0.08)	0.84 (0.08)	0.88 (0.10)
4th-5th	1.12 (0.18)	0.11 (0.05)	0.61 (0.07)	0.54 (0.08)	0.91 (0.10)
U	0.16 (0.08)	0.00 (0.00)	0.47 (0.05)	0.46 (0.06)	0.82 (0.09)

table reports average values (standard errors in parenthesis)

Table 1: Descriptive statistics of developmental variables.

As we can see, mistakes in the transitive reasoning task and attention trials steadily decline with age, and there is a sharp drop in the adult population. The decline is more gentle when we look at discrepancies between Choice and Ranking task for Goods and Social and it is non-existent for Risk. This reinforces the result highlighted in different parts of the paper that the Risk domain follows a different developmental trajectory from the Goods and Social domains.

To better disentangle the effect of the different variables, we ran OLS regressions on the full sample (full) to assess the explanatory power of each variable in each domain. To be precise, we used the number of TV of each participant as the dependent variable and we included *TR mistake*, *Attention* and *Discrepancy* as independent variables. We also included dummy variables for each children's age group (*K-1, 2-3, 4-5*), with the control undergraduate population (U) as the omitted category. Since, by definition, simple policy users do not commit transitivity violations, for the Social and Risk domains we also ran the analysis without those subjects. The results for Goods (columns 1-2-3), Social (columns 4-5-6-7) and Risk (columns 8-9-10-11) domains are reported in Table 2. Columns 7 and 11 correspond to regressions based on the subset of subjects who did not use simple policies (restricted).

	Goods full (1)	*Goods* full (2)	*Goods* full (3)	*Social* full (4)	*Social* full (5)	*Social* full (6)	*Social* restricted (7)	*Risk* full (8)	*Risk* full (9)	*Risk* full (10)	*Risk* restricted (11)
TR mistake	0.39**	0.35*	0.19	0.12	0.13	0.02	-0.10	0.27	-0.11	0.03	0.05
Attention	–	2.58***	2.57***	–	1.37***	1.33***	1.36***	–	1.68**	1.80***	2.02***
Discrepancy	–	2.72***	2.40***	–	1.69***	1.65***	1.35***	–	2.21***	2.24***	2.00***
K-1	3.05***	–	1.38**	1.49***	–	0.78*	1.86***	-0.36	–	-0.95	-0.41
2-3	1.38**	–	0.22	0.74*	–	-0.03	0.33	0.07	–	-0.15	0.64
4-5	0.15	–	-0.28	0.22	–	0.03	0.31	0.77	–	0.60	1.00
Constant	0.69*	-0.47	-0.40	0.53*	-0.17	-0.21	0.01	2.39***	0.60	0.59	0.85*
# obs	185	185	185	184	184	184	125	183	183	183	128
R^2	0.242	0.447	0.485	0.103	0.330	0.364	0.351	0.022	0.380	0.416	0.380

Significance levels: * = 0.05, ** = 0.01, *** = 0.001.

Table 2: OLS regression of TV in Goods, Social and Risk domains

Mistakes in transitive reasoning were not associated with TV in the Social and Risk domains. They were correlated with TV in the Goods domain, but the effect decreased as we controlled for other explanatory variables (2), and disappeared when we also added age group dummies (3). By contrast, performance in attention trials and the ability to make choices consistent with explicit rankings were highly significant in all three domains, whether age dummies were used or not, and whether policy users were included in the sample or not.

The significant effect of performance in attention trials indicates that attentional control played an important role in choice consistency. On the other hand, the transitive reasoning task was not a predictor, suggesting that logical thinking per se is not a key determinant of choice transitivity. This was a surprising result. We expected that the ability to realize that A is greater than C whenever A is greater than B and B is greater than C would develop concurrently with the ability to choose A over C whenever A is chosen over B and B is chosen over C. This was not the case in our data. While more research on the subject is needed, this preliminary finding indicates that consistent reasoning and consistent choices are different abilities, one related to logical, abstract thinking and the other to expression of preferences. As further discussed in section 6.7, it is possible that these calculations are performed in different brain areas.

Finally and notably, participants whose decisions in the Choice tasks were more in line with the ordering displayed in the Ranking tasks also had fewer TV. This effect *is not reminiscent of any known paradigm.* It implies that the gradual and differential development of choice consistency across domains is not entirely explained by changes in attention, centration and logical reasoning. As children learn to know what they like best and least, both their pairwise choices and their explicit rankings become more consistent with each other.[26]

Result 4 *We found partial support for Hypothesis 4: choice consistency is explained by centration and attention control but it is not explained by the ability to perform logical (transitive) reasoning. It is also explained by self-knowledge of preferences, which develops gradually and differentially across domains.*

6.6 Other analyses

We conducted several other statistical analyses and robustness checks. They are briefly described here. Further details can be found in Appendix A.

Other measures of inconsistency. We used the revealed ranking derived in section 6.2

[26] We added a gender dummy variable. It did not have a significant effect in any of the regressions of Table 2. In section 6.6, we study in more detail the relationship between TV and demographic variables.

to determine inconsistencies between this ranking and the pairwise choice comparisons, which we call ICR (Inconsistency in Choices given the Revealed ranking). Similarly, we can also use the explicit ranking obtained in the Ranking tasks to determine inconsistencies between the explicit ranking and the pairwise choice comparisons, which we call ICE (Inconsistency in Choices given the Explicit ranking). In appendix A1, we show that TV in each age group and domain is highly correlated with ICR and ICE. We conclude that while TV is, in our view, the purest way to measure choice (in)consistencies, our basic conclusions are robust to other measures as well.

Reaction time and indifference responses. We checked whether age and domain had an effect on the time it took to make a choice and the likelihood of subjects to be indifferent among alternatives. The information is compiled in Appendix A2. Indifferences are infrequent in general. They are more prevalent in younger children and in the Goods domain. Reaction times are typically longer in trials involved in transitivity violations, suggesting that participants who are more confused, take longer to choose, and end up contradicting their choices more often.

Evolution of preferences. We used the choices in the Social and Risk domains to study the evolution of preferences with age. This analysis is performed in Appendix A3. We notice that social preferences change over time in a way consistent with the literature (Fehr et al., 2008) and risk preferences change from some participants maximizing quantity to many participants being close to maximize expected value. This result has crucial implications. Indeed, a main finding of the paper is the differential learning trajectory over domains: participants learn to know what they like most in the Social domain whereas they learn to know what they like least in the Risk domain. One could argue that this result is simply due to the fact that, despite our attempts to have reasonably balanced options, there is a clear best option in Social and a clear worst option in Risk. Such explanation is at odds with the fact that, in both domains, the most and least preferred options are different across subjects and change significantly with age.

Demographics. Given that violations covaries across domains, we can explore the relationship between TV and demographic variables. In Appendix A4 we present OLS regressions treating TV in each domain as the variable to be explained by a set of demographic characteristics. Those include age group, gender, number of younger siblings and number of older siblings. We found that only the age group predicted TV. Moving from one age group to the next was associated with a decrease in the number of TV in the Goods and Social domains but not in the Risk domain.

6.7 Summary and discussion

The main finding of this section is that logical transitive abilities develop in parallel to transitive choice but the two are not directly related. Transitive decision-making relies on attention control and to self-knowledge of preferences. As we grow, we learn what we like and what we dislike and we become able to draw our choices from explicit preferences. The fact that inconsistency is not associated with the ability to reason transitively suggests that economic choice requires the involvement of different brain regions compared to logical reasoning. Our findings also suggest that two processes interact to produce consistent choice: a 'value' system capable of drawing from explicit preferences and an 'attentional' system that assists the value system. These findings are compatible with current findings in neuroscience in adults. The representation of economic value is associated with regions of the ventromedial prefrontal cortex (Levy and Glimcher, 2012) while logical reasoning involves parietal regions (Hinton, Dymond, Von Hecker, and Evans, 2010). Attentional control involves regions closely related to executive functions and working memory, the latter being known to interact with the value regions during value based decisions (Rudorf and Hare, 2014; Saraiva and Marshall, 2015). All these structures are also known to develop gradually (Gogtay, Giedd, Lusk, Hayashi, Greenstein, Vaituzis, Nugent, Herman, Clasen, Toga, et al., 2004) during childhood and adolescence, which explains the changes in transitive violations counts in our window of observation.

7 Conclusion

We have investigated the developmental trajectories of transitive decision-making in the Goods, Social, and Risk domains in children from Kindergarten to 5th grade and compared consistency levels with adult-level performance. The results obtained for participants in the adult group were consistent with the typical findings in the choice consistency literature. We have reported evidence that transitivity in choices develops gradually, but differentially across domains. The Goods domain closely follows the expected template, with significant improvements with age, high consistency in 4th-5th graders and more mistakes for options ranked in the intermediate range. In the Social domain, a fraction of participants –especially among the youngest ones– utilizes rules of behavior consistent with centration, where each attribute is evaluated sequentially. Among children who do not use such simple rules, we observe the same trajectory as in the Goods domain, suggesting that centration conceals their underdeveloped decision-making system. In the Risk domain, participants are not nearly as consistent as in the other two domains. As in the Social domain, a fraction of participants use simple rules. However, all children who do not use such policies perform at the same level, resulting in no improvement over time.

Our design allows us to identify different learning trajectories across domains. In the Social domain, children learn to pick their most-preferred option. In the Risk domain, children learn to avoid their least-preferred option. In the Goods domain, children learn both. These asymmetries cannot be explained by the existence of an obvious best option in the Social domain and/or an obvious worst option in the Risk domain, since the most and least favorite options change significantly across age groups and across individuals. It may be that, depending on their social orientation and risk tolerance, children can only easily pinpoint their most liked option in the Social domain and their least liked option in the Risk domain.

The analysis also shows that choice transitivity is not related to the development of transitive reasoning and it is only partially explained by the development of attentional control. Instead, it is supported by the improvement of self-knowledge of preferences, which follows a different path in each domain. This effect is further diagnosed by the strong correlation between transitivity violations and inconsistencies between choices and rankings: inconsistent subjects are those who value options differently in different contexts.

These findings can be discussed in the context of our current understanding of how value based decisions are implemented in the adult brain and how these regions are developing over childhood. Indeed, economic decision-making depends critically on the interplay between a value system and an attention system. The functions of these two systems match the two important paradigms identified in our study: self-knowledge of preferences allows to value options, and attentional control allows to devote attention when making close or ambiguous comparisons. Both systems are known to develop gradually, attentional control being part of a set of functions that mature last. Current knowledge is consistent with the observed development in the Goods domain. Yet, we have identified interesting asymmetries across domains that are not reminiscent of any finding and should be further investigated. It is plausible that the way children learn about social and risky options impacts the way they remember them and form value around them.

We would like to conclude by pointing out that choice consistency is the most fundamental hypothesis of economic modeling. Understanding how this ability develops is critical to identify what promotes it. There is growing evidence that the behavior of some populations does not necessarily satisfy choice consistency. As pointed in this article, young children do not make consistent choices, older children learn to make consistent choices in some domains only, and consistency in the risk domain may be a futile assumption until adulthood. Other studies have reached similar findings in older adults (Brocas et al., 2019) and in patients (Camille, Griffiths, Vo, Fellows, and Kable, 2011). The underlying common elements in all these populations are an impaired / underdeveloped value system or an impaired / underdeveloped attention system. This suggests that consistent

decision-making stems from the good functioning of these two systems. More generally, it poses a new challenge to model the preferences and choices of populations exhibiting deficiencies in these systems, which include older adults, addicts and patients suffering from behavioral disorders.

References

James Andreoni and John Miller. Giving According to GARP : An Experimental Test of the Consistency of Preferences for Altruism. *Econometrica*, 70(2):737–753, 2002.

Duncan E. Astle and Gaia Scerif. Using developmental cognitive neuroscience to study behavioral and attentional control. *Developmental Psychobiology*, 51(2):107–118, 2008.

Jori Barash, Isabelle Brocas, Juan D Carrillo, and Niree Kodaverdian. Heuristic to bayesian: The evolution of reasoning from childhood to adulthood. *Journal of Economic Behavior & Organization*, 159:305–322, 2019.

Raymond C Battalio, John H Kagel, Robin C Winkler, Edwin B Fisher, Robert L Basmann, and Leonard Krasner. A test of consumer demand theory using observations of individual consumer purchases. *Economic Inquiry*, 11(4):411, 1973.

Rafal Bogacz, Eric-Jan Wagenmakers, Birte U Forstmann, and Sander Nieuwenhuis. The neural basis of the speed–accuracy tradeoff. *Trends in neurosciences*, 33(1):10–16, 2010.

Samantha Bouwmeester and Klaas Sijtsma. Constructing a transitive reasoning test for 6- to 13-year-old children. *European Journal of Psychological Assessment*, 22(4):225–232, 2006.

Samantha Bouwmeester, Jeroen K Vermunt, and Klaas Sijtsma. Development and individual differences in transitive reasoning: A fuzzy trace theory approach. *Developmental Review*, 27(1):41–74, 2007.

Isabelle Brocas and Juan D. Carrillo. The development of other-regarding behavior and social strategic ignorance from childhood to adulthood. Working Paper, 2018a.

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

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

Isabelle Brocas, Juan D Carrillo, T Dalton Combs, and Niree Kodaverdian. Consistency in simple vs. complex choices by younger and older adults. *Journal of Economic Behavior & Organization*, 157:580–601, 2019.

Karin C Brocki and Gunilla Bohlin. Executive functions in children aged 6 to 13: A dimensional and developmental study. *Developmental neuropsychology*, 26(2):571–593, 2004.

Nathalie Camille, Cathryn A Griffiths, Khoi Vo, Lesley K Fellows, and Joseph W Kable. Ventromedial frontal lobe damage disrupts value maximization in humans. *Journal of Neuroscience*, 31(20):7527–7532, 2011.

Gary Charness, Dan Levin, and David Schmeidler. A generalized winner's curse: An experimental investigation of complexity and adverse selection. Working Paper, 2014.

Syngjoo Choi, Raymond Fisman, Douglas Gale, and Shachar Kariv. Consistency and Heterogeneity of Individual Behavior under Uncertainty. *The American Economic Review*, 97(5):1921–1938, 2007.

J.C. Cox. On Testing the Utility Hypothesis. *The Economic Journal*, 107(443):1054–1078, 1997.

W. Crain. *Theories of Development: Concepts and Applications*. Psychology Press, 2015.

Matthew C. Davidson, Dima Amso, Loren Cruess Anderson, and Adele Diamond. Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, 44 (11):2037–2078, 2006.

Margaret Donaldson. Conservation - What is the question? *British Journal of Psychology*, 23(2):199, 1982.

Richard A Fabes and Nancy Eisenberg. Meta-analyses of age and sex differences in children's and adolescents' prosocial behavior. *Handbook of Child Psychology,*, 3, 1998.

Ernst Fehr, Helen Bernhard, and Bettina Rockenbach. Egalatarianism in young children. *Nature*, 454(28):1079–1084, 2008.

Raymond Fisman, Shachar Kariv, and Daniel Markovits. Individual Preferences for Giving. *The American Economic Review*, 97(5):1858–1876, 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):177–190, 2004.

Nitin Gogtay, Jay N Giedd, Leslie Lusk, Kiralee M Hayashi, Deanna Greenstein, A Catherine Vaituzis, Tom F Nugent, David H Herman, Liv S Clasen, Arthur W Toga, et al. Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National academy of Sciences of the United States of America*, 101(21):8174–8179, 2004.

Joshua I Gold and Michael N Shadlen. Neural computations that underlie decisions about sensory stimuli. *Trends in cognitive sciences*, 5(1):10–16, 2001.

William T Harbaugh and Kate Krause. Children's Contributions in Public Good Experiments: The Development of Altruistic and Free-riding Behaviors. *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.

William T Harbaugh, Kate Krause, and Lise Vesterlund. Risk Attitudes of Children and Adults : Choices Over Small and Large Probability Gains and Losses. *Experimental Economics*, 5(1):53–84, 2002.

Margaret Harris and George Butterworth. *Developmental psychology: A student's handbook*. Psychology Press, 2012.

EC Hinton, S Dymond, Ulrich Von Hecker, and Christopher John Evans. Neural correlates of relational reasoning and the symbolic distance effect: Involvement of parietal cortex. *Neuroscience*, 168(1):138–148, 2010.

Ian Krajbich, Dingchao Lu, Colin Camerer, and Antonio Rangel. The attentional drift-diffusion model extends to simple purchasing decisions. *Frontiers in psychology*, 3:193, 2012.

Dino J. Levy and Paul W. Glimcher. The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, 22(6):1027–1038, 2012.

John A List and Daniel L Millimet. The Market: Catalyst for Rationality and Filter of Irrationality. *The BE Journal of Economic Analysis & Policy*, 8(1):Article 47, 2008.

Graham Loomes, Chris Starmer, and Robert Sugden. Observing violations of transitivity by experimental methods. *Econometrica*, pages 425–439, 1991.

Aurelio Mattei. Full-scale real tests of consumer behavior using experimental data. *Journal of Economic Behavior & Organization*, 43(4):487–497, 2000.

Jean Piaget. La psychologie de l'intelligence. *Revue Philosophique de Louvain*, 46(10): 225–227, 1948.

Jean Piaget. The child's conception of numbers. *Trans. eds C. Gattegno and FM Hodgson, NY: Routledge*, 1952.

Jean Piaget, David Elkind, and Anita Tenzer. *Six psychological studies.* New York, NY: Vintage Books., 1967.

Valerie F Reyna and Susan C Ellis. Fuzzy-trace theory and framing effects in children's risky decision making. *Psychological Science*, 5(5):275–279, 1994.

Sarah Rudorf and Todd A Hare. Interactions between dorsolateral and ventromedial prefrontal cortex underlie context-dependent stimulus valuation in goal-directed choice. *Journal of neuroscience*, 34(48):15988–15996, 2014.

P A Samuelson. A Note on the Pure Theory of Consumer's Behaviour. *Economica*, 5(17): 61–71, 1938.

Ana Carolina Saraiva and Louise Marshall. Dorsolateral–ventromedial prefrontal cortex interactions during value-guided choice: A function of context or difficulty? *Journal of Neuroscience*, 35(13):5087–5088, 2015.

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.

Jan Smedslund. Transitivity of preference patterns as seen by preschool children. *Scandinavian journal of psychology*, 1(1):49–54, 1960.

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.

Hal R Varian. The Nonparametric Approach to Demand Analysis. *Econometrica*, 50(4): 945–973, 1982.

Appendix A: additional statistical analysis

A1. Other measures of choice inconsistencies

Inconsistency in Choices given the Revealed ranking (ICR). We use the implicit ranking described in section 6.2 to construct a new measure of inconsistency, using the following simple rule. Suppose that, according to the tally of the implicit ranking, option B is ranked lower than A. The participant receives a score of 0, if A is chosen over B (showing consistency with the implicit ranking), a score of 0.5 if he expresses indifference between A and B, and a score of 1 if B is chosen over A (showing inconsistency with the implicit ranking). We call this classification error score ICR (Inconsistency in Choices given the Revealed ranking). Notwithstanding the endogeneity problem with this measure (we use choices to extract a revealed ranking then check that ranking against the very choices that were used to create it), it still provides a measure of choice "noisiness." As can be seen by comparing Figure 5 with Figure 16 (left), the age pattern of choice inconsistencies with implicit rankings (ICR) closely resembles that of TV: participants who make more TV are those who are more "noisy" (make more mistakes) around their implicit ranking.

Inconsistency in Choices given the Explicit ranking (ICE). As an additional measure, we evaluate inconsistencies between choices in Choice tasks and the explicit rankings elicited in Ranking tasks. This measure is computed in the same way as for ICR, that is, we check the subject's choices for inconsistencies against their explicit ranking. We called ICE (Inconsistency in Choices given the Explicit ranking) the new classification error score. One advantage of this measure over the previous one is the absence of endogeneity (errors are computed from rankings in two different tasks). The results are illustrated in Figure 16 (right).

In the Goods domain, participants in the **K-1st** and **2nd-3rd** age groups have relatively more difficulty making choices consistent with their explicit rankings compared to older participants (p-values < 0.017 for all comparisons). A similar story holds qualitatively in the Social domain, except that the participants in the **K-1st** and **2nd-3rd**, except for a less gradual development of the ability to choose from an explicit ranking. We found that inconsistencies in the Social domain follow the same trend as inconsistencies in the Goods domain after removing subjects who use simple policies. In the Risk domain, inconsistencies are high and similar across all subjects: they are not able to make choices consistent with their explicit rankings. This result changes when we removed subjects who used simple policies: they become more capable over time to express their explicit rankings in their choices. Nevertheless, the level of inconsistencies remain higher than in the Goods domain for all ages.

Relationship between measures of consistency. Not surprisingly given our previous

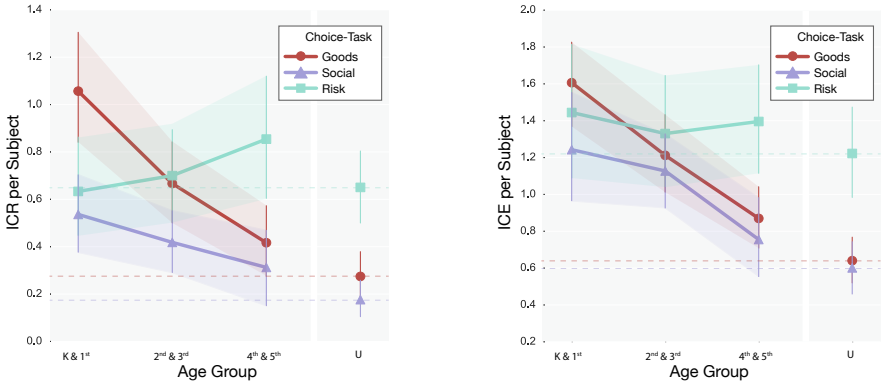

Figure 16: Ranking and choices. Left: Inconsistencies with respect to revealed ranking (ICR). Right: Inconsistencies with respect to explicit ranking (ICE).

findings, our measures of consistency are all highly correlated. Indeed, a high number of transitivity violations (TV) is associated with a high number of inconsistencies between explicit rankings and choices (ICE) in the Goods (Pearson = 0.79), Social (Pearson = 0.69) and Risk (Pearson = 0.72) domains, with all p-values < 0.0001.

Overall, although we feel that TV is the best measure of choice consistency, the main conclusion of Appendix A1 is that the results presented in the main text are robust to different inconsistency measures, such as ICE or ICR: intransitivity is invariably associated with the inability to make choices consistent with rankings, both implicit and explicit, and these have different developmental signatures across domains.

A2. Other analyses of choices

Indifferences. We check whether age and choice domain have an influence on the likelihood to be indifferent in a pairwise comparison. The results are reported in Table 3. In the Goods domain, the number of indifferent choices decreases over time (unpaired t-test, p-values < 0.05 between **K-1st** and **4th-5th**, between **K-1st** and **U**,and between **2nd-3rd** and **U**). Participants are less often indifferent in the Social domain than the Goods domain (paired t-test, p-values < 0.05 for age groups **K-1st** and **U**), and school-age children in the Social domain are more often indifferent than participants in the **U** age group (unpaired t-test, p-values < 0.05). Finally, in the Risk domain we observe no trend in reducing indifferences with age.

At the same time, triplets involving indifferent choices are less likely to result in TV in all domains for the whole sample (t-tests, p-values < 0.0001), as well as for each age group (p-values < 0.0001).

	Goods	Social	Risk
K-1st	3.81 (0.54)	2.04 (0.47)	1.53 (0.59)
2nd-3rd	3.09 (0.47)	2.47 (0.54)	1.31 (0.45)
4th-5th	2.19 (0.42)	2.50 (0.60)	1.94 (0.53)
U	1.86 (0.23)	1.12 (0.28)	1.49 (0.35)

(standard errors in parenthesis)

Table 3: Average number of indifferences.

Reaction times and choices. We found that reaction times in trials involving violations are longer compared to trials involving no violations. This is true in all age groups and in all domains (KS test, p-value = 0.015 for **K-1st** in the Risk domain, p-value = 0.010 for **2nd-3rd** in the Risk domain and p-value < 0.0001 for all other age groups and domains), suggesting that participants who are more confused, take longer to choose an end up contradicting their choices. We also found that reaction times are longer when participants are indifferent compared to when they are not (KS test, p-value < 0.0001 for age groups above **2nd-3rd**). This is consistent with the idea that choices are more difficult to make when options are close in value. However, among trials in which the indifference button is pressed, those involved in a violation take slightly less time than those not involved in a violation (KS test, p-value = 0.053). This is consistent with the results obtained regarding indifferences: participants spend more time on choices where they end up pressing the indifference button and avoid a violation. Last, simple policy users are faster than the other subjects (KS test, p-value < 0.0001), indicating that their choice process is simpler. These additional measures confirm and complement the idea that TV are associated with confusion about one's preferences, which decrease with age.

A3. Evolution of preferences

Evolution of preferences in the Social domain. The most common simple policy employed by our participants is to maximize the reward for self, but its predominance changes over time. Indeed, the preference for giving seems to be developing: small children tend to maximize their own reward systematically and, other things being equal, they also prefer smaller rewards for others. Older participants however select larger rewards for others. We do not find any evidence that participants maximize the sum of payoffs of both participants. However, we found that they come closer to that policy with age, with participants in the **4th-5th** age group being similarly close to it as participants in the **U** group.

The most and least favorite options, as revealed by implicit rankings, also changes over time. Indeed, for all ages, (4,0) and (3,3) are the most popular, but the frequency of (4,0)

decreases while the frequency of (3,3) increases with age. Similarly, (0,4) and (0,5) are the least popular but the frequency of (0,5) decreases while the frequency of (0,4) increases with age. The results are summarized in Table 4.[27]

	Most favorite		Least favorite	
	(4,0)	(3,3)	(0,4)	(0,5)
K-1st	0.65	0.33	0.43	0.77
2nd-3rd	0.40	0.58	0.31	0.78
4th-5th	0.50	0.50	0.50	0.72
U	0.33	0.73	0.80	0.30

Table 4: Most and least favorite options in Social domain derived from implicit rankings

Our Social-Choice task is also rich enough to permit a study of the evolution of other regarding preferences. In particular, it is possible to study whether participants are prosocial (choose (3,3) over (3,1)), willing to share (choose (3,1) over (4,0)) or envious (choose (0,4) over (0,5)). In other words, we can conduct a similar analysis as in Fehr et al. (2008), and determine a type for each subject as a function of the decisions in these three pairs of choices. We define the same five types as in Fehr et al. (2008): "strongly egalitarian" (choices (3,3); (3,1); (0.4)), "weakly egalitarian" (choices (3,3); (4,0); (0,4)), "strongly generous" (choices (3,3); (3,1); (0,5)), "weakly generous" (choices (3,3); (4,0); (0,5)), and "spiteful" (choices (3,1); (4,0); (0,4)). To make it comparable with the literature, we exclude subjects who express one or more indifferences. In Figure 17, we report the proportion of subjects in each of these five categories (as in Fehr et al. (2008), the remaining subjects are those who do not belong to any of them).

Notice that results in these two studies are not directly comparable: ages overlap, and reward and options are different (e.g., choosing between (1,1) and (1,2) captures a different aspect of envy than choosing between (0,4) and (0,5)). Furthermore, our study is not designed specifically to test for social preferences. With this caveat in mind, if we compare 5y-6y with **K-1st** and 7y-8y with **2nd-3rd** we still notice that outcomes are similar across studies. Consistent with the centration hypothesis, many young children are spiteful. As they grow, they develop some integrative reasoning and become more egalitarian. Our oldest participants are predominantly generous.

Evolution of preferences in the Risk domain. The most common simple policy employed by our participants is to maximize reward but, as in the Social domain, its predominance changes over time. More than 20% of participants in the **K-1st** and **2nd-3rd** age groups

[27] Some percentages exceed 100 because some subjects put more than one option in the same ranking, so we count these options twice.

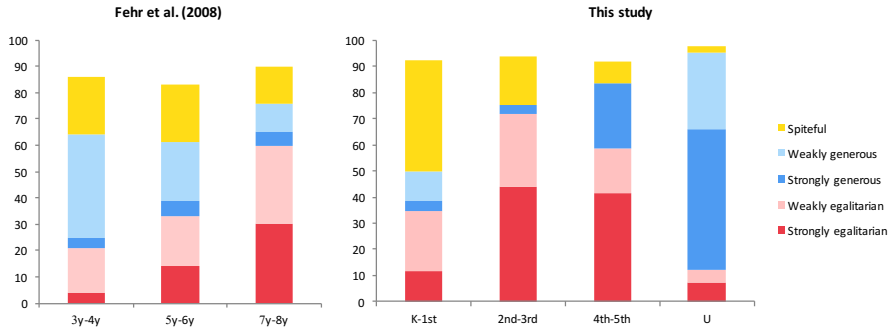

Figure 17: Evolution of other-regarding preferences.

use it against less than 10% in the **4th-5th** and **U** age groups. We choose maximization of expected value, E(V), as a template of integrative reasoning, but strictly speaking no participant maximizes E(V). Only two subjects are one step away from that integrative policy and they both have some TV. For each participant, we counted the number of choices that maximize E(V) and averaged this count across participants in each Choice task and each age group. We found that participants in the **K-1st**, **2nd-3rd** and **4th-5th** age groups make significantly fewer choices consistent with the maximization of E(V) compared to participants in the **U** age group. In particular, after removing simple policy users, each group of children uses policies that are farther away from E(V) compared to participants in the **U** age group (t-test, p-values < 0.005).

When looking at the favorite option revealed by implicit rankings, we found that children transition gradually from the option involving the largest quantity $(12, 12.5\%)$ to the option exhibiting the largest expected value $(5, 50\%)$, as depicted in Table 5. Interestingly, option $(12, 12.5\%)$ is ranked first by many and, at the same time, last by others.

	(12,12.5%)	**(5,50%)**
K-1st	0.60	0.13
2nd-3rd	0.44	0.35
4th-5th	0.28	0.53
U	0.16	0.82

Table 5: Favorite option in the Risk domain derived from implicit rankings

Taken together, the results in Social and Risk show that behavior changes from (some) choices consistent with very simple policies to choices resulting from trade-offs and integrative thinking. The centration effect observed in some of our young participants makes

them appear selfish and consistent in the Social domain and risk-loving and consistent in the Risk domain. These attitudes gradually change with age. As discussed in the text, this evolution in preferences suggests that the options selected were comparable, that is, there was not a clear best option in Social or a clear worst option in Risk that would trivially explain the inconsistency patterns highlighted in Results 3b and 3c.

A4. Relationship between TV and demographic variables

For each domain, we ran OLS regressions to study TV as a function of three demographic variables: *Gender* (female = 1), number of older siblings (*Older-s*) and number of younger siblings (*Younger-s*). In each regression, we include data from only two consecutive age groups: **K-1st** and **2nd-3rd** (regressions (1), (4) and (7)), **2nd-3rd** and **4th-5th** (regressions (2), (5) and (8)), **4th-5th** and **U** (regressions (3), (6) and (9)). We also add an age-group dummy variable to study the effect of moving from one age-group to the next. Finally, we include a dummy for simple policy users (*Policy*). The results are summarized in Table 6.

	Goods (1) K-1/2-3	*Goods* (2) 2-3/4-5	*Goods* (3) 4-5/U	*Social* (4) K-1/2-3	*Social* (5) 2-3/4-5	*Social* (6) 4-5/U	*Risk* (7) K-1/2-3	*Risk* (8) 2-3/4-5	*Risk* (9) 4-5/U
Gender	0.39	1.13	0.35	-0.06	0.66	0.23	0.74	1.07	0.54
Older-s	-0.04	-0.08	-0.10	0.16	0.04	-0.27	-0.32	-0.53	0.21
Younger-s	-0.90	-0.40	0.24	0.15	-0.05	-0.22	0.17	0.06	-0.12
2-3	-1.91**	–	–	-1.24**	–	–	-0.08	–	–
4-5	–	-1.34*	–	–	-0.36	–	–	-0.28	–
U	–	–	-0.55	–	–	-0.41	–	–	-1.51*
Policy	–	–	–	-2.58**	-1.45**	-1.04**	-4.26**	-4.19**	-3.14**
Constant	4.79**	2.39**	1.00**	3.06**	1.49**	1.43**	4.40**	4.35**	3.87**
# obs	102	87	83	101	87	83	100	87	83
R^2	0.10	0.10	0.08	0.25	0.21	0.22	0.46	0.38	0.18

Significance levels: * = 0.05, ** = 0.01.

Table 6: OLS regression of TV on demographic variables

Gender, number of older siblings and number of younger siblings have no explanatory power in any regression, while age explains transitivity violations across some children age-groups in Goods and Social. The results do not change if we remove simple policy users as an explanatory variable.