Cognitive (Ir)reflection: New Experimental Evidence*

Carlos Cueva† Iñigo Iturbe-Ormaetxe† Esther Mata-Pérez†
Giovanni Ponti‡ Marcello Sartarelli† Haihan Yu†
Vita Zhukova†

Abstract

We study how cognitive abilities correlate with behavioral choices by collecting evidence from almost 1,200 subjects across eight experimental projects concerning a wide variety of tasks, including some classic risk and social preference elicitation protocols. The Cognitive Reflection Test (CRT) has been administered to all our experimental subjects, which makes our dataset one of the largest in the literature. We partition our subject pool into three groups depending on their CRT performance. *Reflective* subjects are those answering at least two of the three CRT questions correctly. *Impulsive* subjects are those who are unable to suppress the instinctive impulse to follow the intuitive -although incorrect- answer in at least two 2 questions. The remaining subjects form a *residual group*. We find that females score significantly less than males in the CRT and that, in their wrong answers, impulsive ones are observed more frequently. The 2D-4D ratio, which is higher for females, is correlated negatively with subjects' CRT score. We also find that differences in risk attitudes across CRT groups crucially depend on the elicitation task. Finally, impulsive subjects have higher social (inequity-averse) concerns, while reflective subjects are more likely to satisfy basic consistency requirements in lottery choices.

JEL Classification: C91, D81, J16

Keywords: behavioral economics, cognitive reflection, gender effects, experiments.

---

*We thank Enrica Carbone, Xavier Del Pozo, Daniela Di Cagno, Arianna Galliera, Glenn Harrison, Raffaele Miniaci, Ismael Rodriguez-Lara, Iryna Sikora and Josefa Tomás for letting us using data from projects carried out with their direct involvement. The usual disclaimers apply. Financial support from the Spanish Ministries of Education and Science and Economics and Competitiveness (SEJ 2007-62656, ECO2011-29230, ECO2012-34928 and ECO2013-43119), Universidad de Alicante (GRE 13-04), MIUR (PRIN 20103S5RN3_002), Generalitat Valenciana (Research Projects Gruposo3/086 and PROMETEO/2013/037) and Instituto Valenciano de Investigaciones Económicas (IVIE) is gratefully acknowledged.

†Universidad de Alicante.

‡Corresponding author. Universidad de Alicante and LUISS Guido Carli Roma. Dipartimento di Economia e Finanza, LUISS Guido Carli. Viale Romania, 32. 00185 Roma. Email: gponti@luiss.it

1 Introduction

There is a growing literature that studies the link between various aspects of socio-economic behavior, such as risk, time, or social preferences, and proxies of cognitive ability of various formats. These measures vary from school and college performance, such as the Grade Point Average (GPA, Kirby et al., 2005), college entry standardized test scores, such as GRE or SAT (Dohmen et al., 2010; Chen et al., 2013), up to more customized protocols, from the classic IQ test (Borghans et al., 2008b) to the Wonderlic test, aimed at assessing problem-solving ability (Ben-Ner et al., 2004).[1] All these contributions stress the importance of *individual heterogeneity*, with specific reference to cognitive abilities, as a fundamental factor to understand and predict individual and social behavior.

Cognitive ability is also a fundamental component of all theories that advocate a dual and parallel cognitive deliberation process (Evans, 1984; Kahneman, 2011): one ("System 1", or intuitive, heuristic...) fast, automatic, associated with a low cognitive load, the other ("System 2", or controlled, analytic...) more cognitively demanding. The Cognitive Reflection Test (CRT hereafter, Frederick, 2005) illustrates the interaction between these two cognitive processes. It is a simple test of a quantitative nature especially designed to elicit the "predominant cognitive system at work", either 1 or 2, in respondents' reasoning:

CRT1. A bat and a ball cost 1.10 dollars. The bat costs 1.00 dollars more than the ball. How much does the ball cost? (**Correct answer: 5 cents**).

CRT2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? (**Correct answer: 5 minutes**).

CRT3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? (**Correct answer: 47 days**).

The beauty of the test is that, to each question, is associated an immediate, "impulsive", answer (10, 100 and 24, respectively) that, although incorrect, may be

[1] The Wonderlic test consists of 50 questions in the areas of math, vocabulary, and reasoning and its score is positively correlated with various measures of intelligence (Hawkins et al., 1990).

selected by those subjects who do not think carefully enough. As Frederick (2005, p. 27) puts it, "*. . . the three items on the CRT are easy in the sense that their solution is easily understood when explained, yet reaching the correct answer often requires the suppression of an erroneous answer that springs "impulsively" to mind*".

Frederick (2005) shows that CRT performance significantly correlates with risk and time preferences: more reflective subjects are, on average, less risk-averse and more patient. Recent studies also document that the CRT is associated with subjects' gender-specific exposure to testosterone (Bosch-Domènech et al., 2014). In addition, it helps to explain some classic biases in behavioral finance, such as the so-called "base rate fallacy" (Bergman et al., 2010; Hoppe and Kusterer, 2011; Oechssler et al., 2009; Alos-Ferrer and Hügelschäfer, 2015; Noussair et al., 2014; Kiss et al., 2015; Insler et al., 2015).

The CRT has also gained attention for the fact that, contrary to other proxies of cognitive abilities such as the SAT or the Wonderlic Test, *females score significantly less than males*. This stylized fact has been established in a wide variety of studies (Frederick, 2005; Hoppe and Kusterer, 2011; Oechssler et al., 2009) and is also confirmed by the evidence reported in this paper.

It may be worth highlighting that the CRT provides not only a measure of cognitive abilities, but also of *impulsiveness* and, possibly, other individual unobservable characteristics. For instance, the number of correct answers in the CRT has been shown to be positively correlated with numerical literacy, mathematical skills, and various psychological dimensions (Morsanyi et al., 2014; Toplak et al., 2011; Borghans et al., 2008a). This means that the CRT alone cannot reveal the cognitive and psychological mechanisms underlying individual heterogeneity in economic behavior. For instance, it is possible that subjects performing high in the CRT are closer to risk neutrality because they are less impulsive or because they better understand the decision problems at stake. This is why, in this paper, we look closely at the relationship between CRT performance and physiological, psychological and socio-demographic characteristics (Section 3). In addition, we also relate CRT scores to alternative measures of cognitive ability, such as financial literacy and consistency in risky choices (Section 6).

In the last five years, the CRT has been administered to the participants in eight experimental studies, both at LaTEx and CESARE, the experimental labs of the Universidad de Alicante and LUISS Guido Carli in Rome, respectively, for a total of nearly 1,200 observations (see Section 2 for a detailed description). To get directly into the discussion around which this paper is built, Figure 1 reports the distribution of CRT answers of our compound dataset. As Figure 1 shows, in none of the cases the modal response corresponds to the correct answer. Instead, the mode (10, 100 and 24, respectively) is always associated with *"the erroneous answer that springs impulsively to mind"*. In this respect, our evidence is perfectly in line with what is reported in the literature: for all three questions, the impulsive (System-1) responses are much more frequent than the reflective (System-2) ones (Gill and Prowse, 2015).

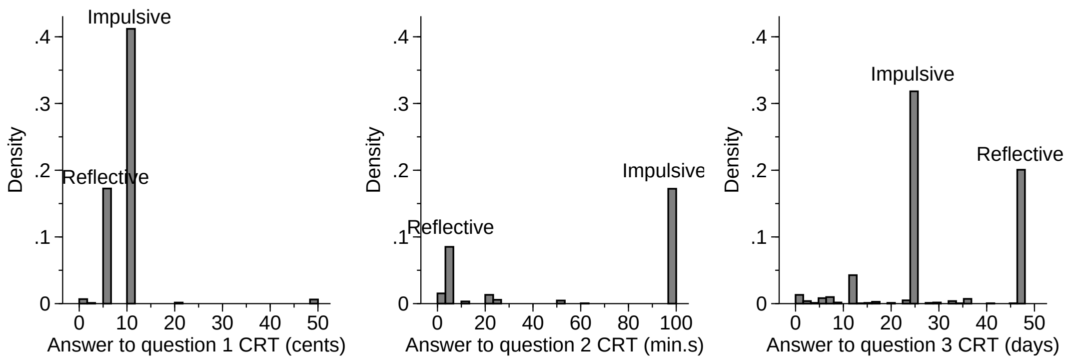


Figure 1: CRT answers distributions.

Figure 1 also shows that the response distribution is not completely polarized between these two answers: there are also alternatives -neither reflective, nor impulsive- that are selected by a non negligible fraction of individuals. These subjects' answers fall short with respect to the dichotomy "reflective-impulsive" along which the discussion on CRT performance has often focused upon (see, e.g., Frederick, 2005; Brañas-Garza et al., 2012; Grimm and Mengel, 2012).

In order to further investigate this issue, this paper puts forward an additional index, labelled as iCRT, which is meant to measure cognitive "impulsiveness" by means of the same three CRT questions:

$$iCRT = 1(CRT1 = 10) + 1(CRT2 = 100) + 1(CRT3 = 24),$$

where 1(.)=1 if condition (.) is satisfied, and 0 otherwise. By analogy with the standard CRT score, an index from 0 to 3 that counts the number of correct answers in the CRT, our *i*CRT is meant to measure the inability to *suppress the erroneous intuitive answer*, which in our view provides as important information as the CRT score in characterizing our subject pool. As our previous discussion suggests, we expect females to have, on average, higher *i*CRT scores, but additional behavioral dimensions need to be explored.

Panel A in Figure 2 reports the distribution of CRT scores disaggregated by gender. The mode is zero for both genders, but the fraction of females who fail the three questions is much higher than the corresponding fraction of males. By the same token, males' average CRT score is significantly higher (1.12 *vs.* 0.58, $p < 0.001$), while the opposite holds for the *i*CRT score (1.46 *vs.* 1.93, $p < 0.001$). However, there is also a significant fraction of subjects (19% of our pool) who score "low" (i.e., not more than 1 correct answer) in both CRT and *i*CRT, thus suggesting that cognitive (ir-)reflection does not seem to fully explain their cognitive processing. These considerations yield the partition of Panel B, where subjects are assigned to one of three categories, depending on whether: i) they scored 2 or more in the CRT ("Reflective"), ii) they scored 2 or more in the *i*CRT ("Impulsive"), or iii) they scored poorly in both tests (≤ 1, "Residual"). As we see from Panel B of Figure 2, while the first two groups have a strong gender component, the latter distributes across genders almost equally.

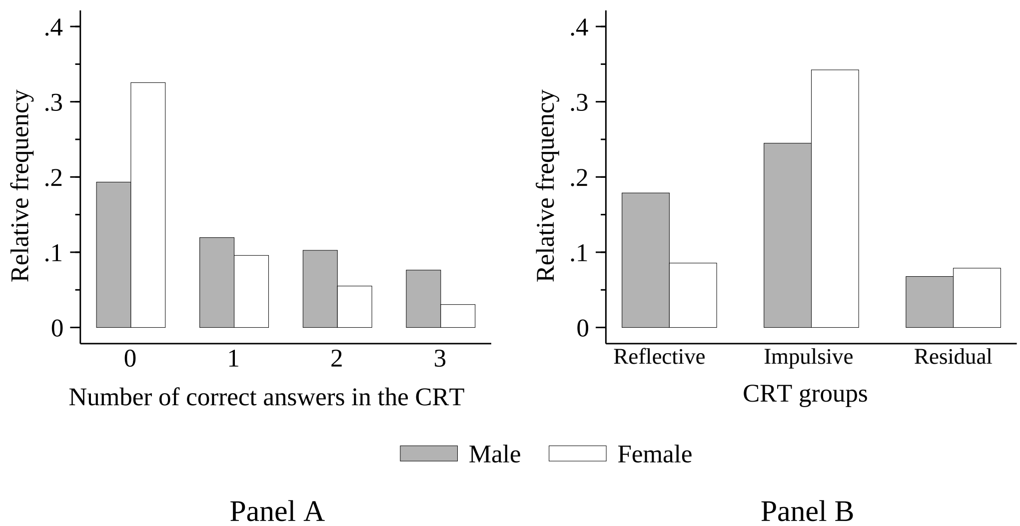


Figure 2: Panel A: CRT score frequencies by gender. Panel B: CRT groups by gender.

The remainder of this paper follows the basic layout of Frederick (2005), in that we provide additional evidence on risk aversion, gender differences, or the relation between CRT and alternative proxies of cognitive ability, around which the original debate on cognitive reflection has been developing over the last 10 years. In addition, we enrich the discussion along less explored dimensions, such as social preferences.

More specifically, Section 2 provides a brief description of the structure of our dataset and the associated experimental projects. Section 3 correlates CRT scores with subjects' observable characteristics grouped into three broad categories: *physiological*, *psychological* and *socio-demographic*. We find that the large gender difference in CRT performance is significant even after including a large number of these individual controls.

Sections 4 and 5 use our behavioral evidence to look into the link between cognitive reflection and risk and social preferences, respectively. As for the former, we show that the negative correlation between CRT performance and risk aversion crucially depends on the elicitation protocol, thus confirming the evidence in Andersson et al. (2013). As for the latter, we find that our CRT partition uncovers novel evidence on the relation between cognitive reflection and social preferences: impulsive subjects have greater (inequity averse) distributional concerns than the other two groups.

In Section 6 we relate CRT performance to alternative measures of cognitive ability. Here we find that reflective subjects are more likely to satisfy some basic "consistency" requirements in their lottery choices and have, on average, higher grades at college (Frederick, 2005). Finally, Section 7 concludes, followed by an Appendix containing supplementary empirical evidence.

2 Data and methods

We collect data from eight experimental studies carried out at the Laboratory of Theoretical and Experimental Economics (LaTEx) of the Universidad de Alicante and the Center for Experimental Studies At Rome East (CESARE) of LUISS Guido Carli in Rome, from 2009 to 2015. The objects of interest include risk and social preferences, mechanism design and behavioral finance. All experimental protocols are also endowed with a computerized debriefing questionnaire.[2]

2.1 Individual characteristics

Table 1 summarizes the structure of our dataset. The behavioral content of the 8 projects is divided into two broad categories: (IND)ividual and (STR)ategic, depending on the nature of the experimental environment. As we shall report in sections 4 and 5, this paper is mainly devoted to establishing a link between cognitive reflection and individual (as opposed to strategic) behavior, the latter being studied elsewhere, or still in progress (see Section 7 for a "sneak preview" of our preliminary results).

Subjects' individual characteristics are grouped into three broad categories: *physiological, psychological* and *socio-demographic*. Subjects took the CRT test, without monetary incentives, within the debriefing questionnaire.[3] Physiological measures include scanned pictures of both hands, from which we compute the second-to-fourth digit ratio (2D:4D hereafter) following the procedure of Neyse and Brañas-Garza (2014).[4] It has been shown that 2D:4D correlates negatively with prenatal exposure to testosterone (Manning et al., 1998). The relationship between 2D:4D and several individual characteristics, such as risk aversion, competitiveness, prosocial preferences, cognitive ability or career choices has been extensively studied in the literature (Apicella et al., 2008; Coates et al., 2009; Sapienza et al., 2009; Pearson and Schipper, 2012; Bosch-Domènech et al., 2014).[5]

As for subjects' psychological characteristics, we use a reduced version of the "Big Five" personality inventory (Benet-Martinez and John, 1998; John and Srivastava, 1999). In its various forms, the Big Five questionnaire is among the most relied-upon measures of personality in psychology (see, e.g., Digman, 1990; John et al., 2008).

[2]All experiments were computerized using z-tree (Fischbacher, 2007). In all projects, the debriefing questionnaire was administered at the end of the experiment, with the exception of Project 6, in which it was administered at the beginning.

[3]The order in which the 3 CRT questions are presented is always the same, as in Frederick (2005).

[4]After scanning participants' hands, digit length was measured with a ruler, whose measurement precision is 0.5 millimeters.

[5]Figure B1 in the Appendix shows the distribution of 2D:4D in our sample. We have also collected self-assessed subjects' height and weight, from which we have derived the associated Body Mass Index (BMI). As it turns out, BMI has never been found a significant factor in all the statistical exercises contained in this paper and, therefore, has been dropped from the set of regressors.

It measures personality according to five broad dimensions, or "traits": *Openness, Conscientiousness, Extraversion, Agreeableness* and *Neuroticism*.[6] The Big Five test has received increasing attention by economists as a useful tool in explaining heterogeneity in individual preferences (Borghans et al., 2009; Daly et al., 2009), academic achievement and labor market performance (Barrick and Mount, 1991; Judge et al., 1999; Heckman and Rubinstein, 2001; Zhao and Seibert, 2006; Heckman et al., 2006; Borghans et al., 2008a; Heckman and LaFontaine, 2010).

Proj.	Reference	Obs.	IND/ STR	Topic	Quest	2D:4D	BIG5	Risk	Soc. pref.s	Fin. lit.
1	Ponti and Carbone (2009)	48	IND	Herding	Yes	No	Yes	MPL	No	No
2	Di Cagno et al. (2014)	192	IND	Risk/soc. preferences	Yes	No	No	N/A	N/A	No
3	Del Pozo et al. (2013)	192	IND	Risk/soc. preferences	Yes	No	No	RLP	Yes	No
4	Ponti et al. (2014b)	336	STR	Entrepreneurship	Yes	Yes	Yes	MPL	Yes	No
5	Ponti et al. (2014a)	192	IND	Risk/Time preferences	Yes	No	No	N/A	N/A	No
6	Ferrara et al. (2015)	32	STR	Public good/sleep depr.	Yes	No	No	RLP	Yes	No
7	Albano et al. (2014)	92	STR	Procurement auctions	Yes	No	No	No	No	No
8	Cueva et al. (2014)	96	STR	Behavioral finance	Yes	Yes	Yes	MPL	No	Yes
Obs.		1,180			1,180	432	480	704	560	96

Table 1: Structure of the meta-dataset

Among the set of socio-demographics, we use *Family education*, a dummy variable that is positive if either parent holds a university degree and *languages*, another dummy variable that is positive if the subject is fluent in more than two languages.[7]

2.2 Behavioral evidence

With regards to the behavioral evidence, this paper focuses especially on *risk* and *social preferences*, which are elicited in 5 and 3 studies of our dataset, respectively.

Risk preferences. Subjects' risk attitudes have been elicited either by means of a Random Lottery Pair protocol (RLP, Projects 3 and 6) or a Multiple Price List protocol (MPL, Projects 1, 4 and 8).[8]

The RLP protocol consists of a sequence of 24 binary choices between lotteries involving four fixed monetary prizes (0, 5, 10 and 15 Euro). Lotteries are selected from Hey and Orme (1994) original design. Our MPL protocol consists of a sequence

[6] See Table B1 in the Appendix for details.

[7] Information on languages spoken is only available for Spanish students. Our study was conducted in a bilingual region of Spain. Thus, we wanted to measure whether a subject was fluent in any other language in addition to Spanish and Catalan.

[8] In the analysis of Section 6 we drop the evidence from Projects 1 and 6 because the former employs hypothetical payoffs and the latter has insufficient observations.

of 21 binary choices. Option A corresponds to a sure payment whose value increases along the sequence from 0 to 1000 pesetas.[9] Option B is constant across the sequence and corresponds to a 50/50 chance to win 1000 pesetas. For both MPL and RLP, one of the binary choices is selected randomly for payment at the end of the experiment.[10]

Social preferences. The data analyzed in this paper are taken from Project 4 and consist of a sequence of 24 distributional decisions borrowed from Cabrales et al. (2010). Individuals are matched in pairs and must choose one out of four options. An option corresponds to a pair of monetary prizes, one for each subject within the pair. Then, one of the two individuals is chosen randomly to be the "dictator", whose decision is implemented for the pair. This is the so-called "Random Dictator" protocol (Harrison and McDaniel, 2008).[11]

3 CRT and individual characteristics

Table 2 reports mean values of individual characteristics for each CRT group. It also provides p-values from Kruskal-Wallis tests whose null hypothesis is that each individual characteristic follows the same distribution across the three CRT groups.[12] As Table 2 shows, subjects belonging to different CRT groups vary significantly with respect to gender, 2D:4D, Neuroticism, Openness and Agreeableness.[13]

3.1 Physiological

We begin by looking at our two physiological measures, gender and 2D:4D. As we know from Figure 2, both CRT scores and groups have a strong gender component, with the exception of the residual group. As a consequence, the distributions of both CRT scores and groups are significantly different across gender (Mann-Whitney U

[9]It is standard practice, for all experiments ran at LaTEx, to use Spanish Pesetas as experimental currency. The reason for this design choice is twofold. First, it mitigates integer problems, compared with other currencies (USD or Euros, for example). On the other hand, although Spanish Pesetas are no longer in use (substituted by the Euro in the year 2002), Spanish people still use Pesetas to express monetary values in their everyday life. In this respect, by using a real (as opposed to an artificial) currency, we avoid the problem of framing the incentive structure of the experiment using a scale (e.g., Experimental Currency) with no cognitive content.

[10]Figure A1 in the Appendix shows the user interfaces of the MPL and RLP protocols.

[11]The user interface for the distributional decisions is shown in Figure A2 in the Appendix.

[12]The Kruskal-Wallis test is a multiple-sample generalization of the Mann-Whitney U-test (Kruskal and Wallis, 1952). Tables B2 and B3 in the Appendix present further mean values and tests disaggregated by gender.

[13]We also consider grades and financial literacy later in the paper (see Section 6).

	Mean			Kruskal-Wallis	N. obs.
	Reflective	Impulsive	Residual	P-value	
Female	0.324	0.583	0.538	<0.001***	1,178
Left hand 2D:4D	0.970	0.981	0.987	0.015**	431
Right hand 2D:4D	0.967	0.978	0.981	0.064*	432
Neuroticism	0.435	0.507	0.483	0.009***	479
Extraversion	0.582	0.608	0.565	0.175	479
Openness	0.725	0.697	0.655	0.009***	479
Agreeableness	0.694	0.685	0.639	0.022**	479
Conscientiousness	0.689	0.688	0.671	0.485	479
N. languages > 2	0.440	0.368	0.387	0.462	432
Family educ.	0.311	0.296	0.387	0.377	432

Table 2: Mean values of individuals' characteristics by CRT groups and p-value of the Kruskal-Wallis test.

test, $p = 0.001$ and Chi-square test, $p < 0.001$, respectively).

Figure 3 plots mean 2D:4D for each CRT score and group. As Figure 3 shows, 2D:4D is lowest for men and women with maximum CRT scores and, consequently, for those subjects belonging to the reflective group. This relationship seems stronger for males: Kruskal-Wallis tests reject the null hypothesis of no difference in left hand 2D:4D across CRT scores and groups for males ($p = 0.034$ and $p = 0.050$, respectively), but not for females ($p = 0.217$ and $p = 0.668$). With respect to right hand 2D:4D, Kruskal-Wallis tests cannot reject the null hypothesis of no difference across CRT scores and groups for both males ($p = 0.096$ and $p = 0.365$, respectively) and females ($p = 0.297$ and $p = 0.494$).

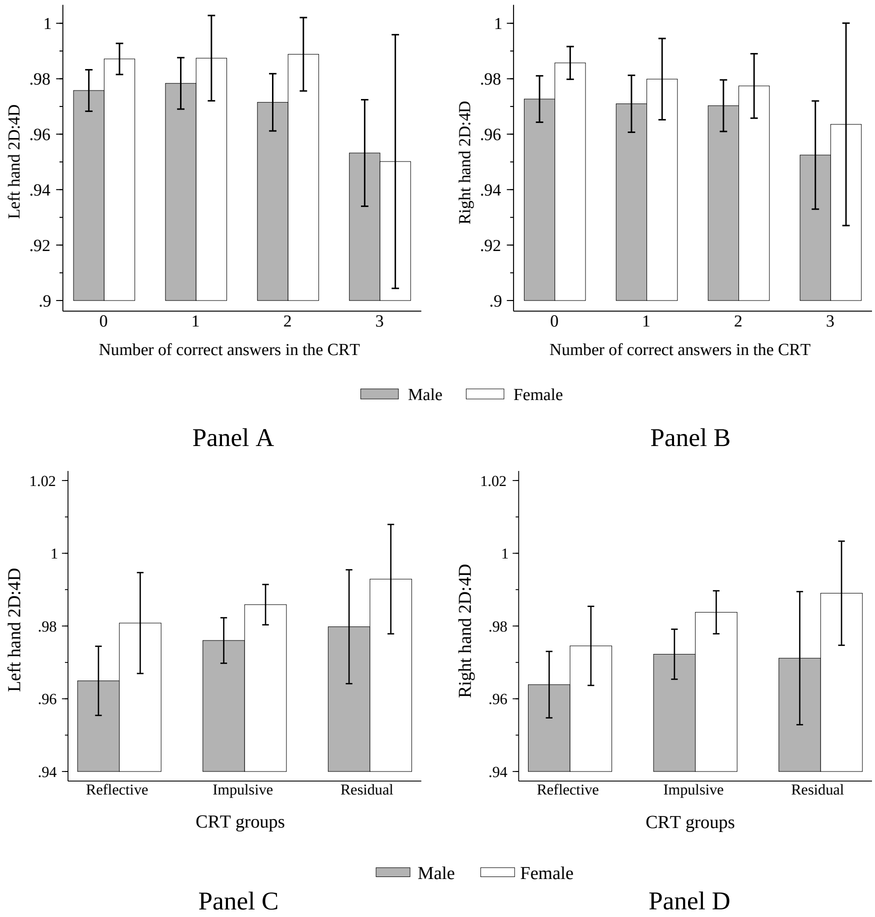


Figure 3: CRT and 2D:4D with 95% confidence intervals. Panel A (B): LH (RH) 2D:4D and CRT. Panel C (D): LH (RH) 2D:4D and CRT groups.

Our finding that males score significantly higher than females in CRT adds further support to the existing literature (Frederick, 2005; Oechssler et al., 2009; Brañas-Garza et al., 2012; Bosch-Domènech et al., 2014). Fewer studies have explored the relationship between 2D:4D and cognitive ability. Brañas-Garza and Rustichini (2011) measure performance in the Raven Progressive Matrices task, a test of abstract reasoning ability and find -consistently with us- a negative and significant correlation between 2D:4D and Raven test scores for males and no significant correlation for females. Bosch-Domènech et al. (2014) study the correlation between 2D:4D and CRT scores and find a negative and significant correlation, particularly with the right hand 2D:4D. However, in contrast with our findings, their correlation is stronger for females.

3.2 Psychological

Table 3 reports the estimated coefficients of some ordered logit regressions in which Big Five scores (interacted with gender) are included in the set of independent variables. As Table 3 shows, in all regressions, Neuroticism and Extraversion are statistically significant.[14] There are no significant interactions between gender and personality traits in our regressions.

	(1)	(2)	(3)	(4)	(5)	(6)
		Left Hand			Right Hand	
2D:4D	-0.181	-0.152	-0.191	-0.220**	-0.190*	-0.148
	(0.111)	(0.113)	(0.148)	(0.104)	(0.105)	(0.135)
Female	-1.117***	-1.028***	-0.973***	-1.111***	-1.020***	-0.939***
	(0.205)	(0.209)	(0.312)	(0.206)	(0.210)	(0.315)
Family education	0.0690	0.0397	-0.0568	0.0652	0.0357	-0.0553
	(0.202)	(0.205)	(0.272)	(0.204)	(0.206)	(0.273)
Languages	0.441**	0.439**	0.606**	0.437**	0.434**	0.613**
	(0.201)	(0.204)	(0.271)	(0.201)	(0.205)	(0.272)
Project 8	-0.228	-0.247	-0.275	-0.253	-0.267	-0.296
	(0.220)	(0.230)	(0.242)	(0.223)	(0.232)	(0.244)
Neuroticism		-0.235**	-0.257*		-0.237**	-0.268**
		(0.100)	(0.131)		(0.0998)	(0.131)
Extraversion		-0.198**	-0.262*		-0.198**	-0.261*
		(0.101)	(0.139)		(0.100)	(0.140)
Openness		0.175	0.110		0.172	0.109
		(0.114)	(0.162)		(0.114)	(0.164)
Agreeableness		-0.0287	-0.0443		-0.0340	-0.0593
		(0.114)	(0.127)		(0.114)	(0.128)
Conscientiousness		-0.0682	-0.108		-0.0636	-0.0966
		(0.106)	(0.151)		(0.106)	(0.150)
Female*2D:4D			0.122			-0.101
			(0.234)			(0.227)
Female*Family education			0.206			0.200
			(0.420)			(0.424)
Female*Languages			-0.382			-0.421
			(0.414)			(0.417)
Female*Neuroticism			0.0502			0.0599
			(0.209)			(0.210)
Female* Extraversion			0.183			0.165
			(0.207)			(0.206)
Female*Openness			0.189			0.163
			(0.242)			(0.242)
Female*Agreeableness			0.0125			0.0529
			(0.249)			(0.255)
Female*Conscientiousness			0.135			0.123
			(0.216)			(0.217)
Observations	431	431	431	432	432	432

Table 3: Ordered Logit estimates of the number of correct answers to CRT. Robust standard errors in parentheses. All explanatory variables except female, languages, family education and project are standardized. *** p <0.01, ** p<0.05, * p<0.1.

Borghans et al. (2008a) examine the impact of personality traits on scores in various cognitive tests, including CRT, in a sample of 128 students. Consistently with us, they find that *Extraversion* is negatively related with the probability of answering correctly. In their data, *Openness* correlates positively with CRT, whereas in

[14]The regressions of Table 3 only consider observations from Projects 4 and 8, since these are the only ones in which we have collected data on the Big Five test.

our regressions the coefficient on *Openness* is also positive, but not significant. Similarly, *Neuroticism* is negatively correlated (although, in their data, the estimated coefficient is not significant).

3.3 Socio-demographic

The regressions of Table 3 include two socio-economic indicators: whether the subject speaks more than two languages and whether at least one parent holds a university degree. Controlling for other variables, speaking more than two languages turns out to be significant, whereas family education is not. Fluency in more than two languages very likely indicates a relatively high socio-economic status in Spain, where the average student is unlikely to be fluent in more than two languages without additional family investment in private education.

3.4 CRT: nature or nurture?

We have used biological, psychological and socio-economic measures as independent variables in our CRT regressions. Our findings that both gender and 2D:4D correlate significantly with CRT, together with those reported in Brañas-Garza and Rustichini (2011) and Bosch-Domènech et al. (2014), lend support to the idea that physiological factors (i.e., *nature*) may affect CRT performance. In contrast, the significant effect of *languages* also suggests that educational investment (i.e., nurture) matters. However, it is difficult to establish a causal relationship here because cognitive ability and intrinsic motivation might themselves affect a subject's ability to learn new languages.

Finally, we also found certain psychological measures to be correlated with CRT. Even though the relative importance of biological and social determinants of personality is less clear, evidence suggests substantial heritability in *Big Five* scores. For instance, twin studies have estimated that genetic influence can account for around 50% of the variance in *Neuroticism* or *Extraversion* (Loehlin, 1992; Jang et al., 1996; Loehlin et al., 1998).

To quantify the effect of our explanatory variables on CRT scores, we predict the probability of having zero correct answers to CRT for different subgroups in our sample.[15] The probability that males answer zero questions correctly is 0.47, controlling for all other covariates, whereas females have a probability of 0.70. Subjects with right hand 2D:4D one standard deviation below average have a probability of 0.56 of having zero correct answers, whereas those with 2D:4D one standard deviation above average have a probability of 0.60. A score one standard deviation above rather than below average in Neuroticism leads to a 9% difference (0.61 and 0.56, respectively). Similarly, a score one standard deviation above rather than below average in Extraversion leads to a 7% difference (0.60 and 0.56, respectively). Finally, subjects speaking more than two languages are 13% less likely to have zero correct answers to CRT than those who do not (0.53 vs 0.62).

In sum, our results highlight the large gender difference in performance in CRT that remains after controlling for other individual variables: females are almost 50% more likely than males to answer all CRT questions wrong. Variations in personality scores or in the digit ratio of two standard deviations led to much more moderate changes in the predicted probability of giving zero correct answers in CRT (7-9%). Finally, our evidence suggests that educational investment (as proxied by the number of languages spoken) could play a more important role than the psychological and physiological characteristics considered here.

4 CRT and risk preferences

We now turn our attention to our behavioral evidence, starting with the analysis on how cognitive reflection relates with risk attitudes. As we already discussed in Section 2, we rely on two different choices formats: RLP and MPL. Contrary to MPL, in RLP lotteries are neither ordered with respect to their associated *profitability* (proxied by the expected return), nor with respect to their associated *risk* (proxied by the variance). Instead, the presentation of each lottery pair is artificially manipulated, precisely to control for possible order effects.

[15] We use the estimates in column (5) of Table 3. Remember from Figure 2 that the modal number of correct answers to CRT is zero for both males and females.

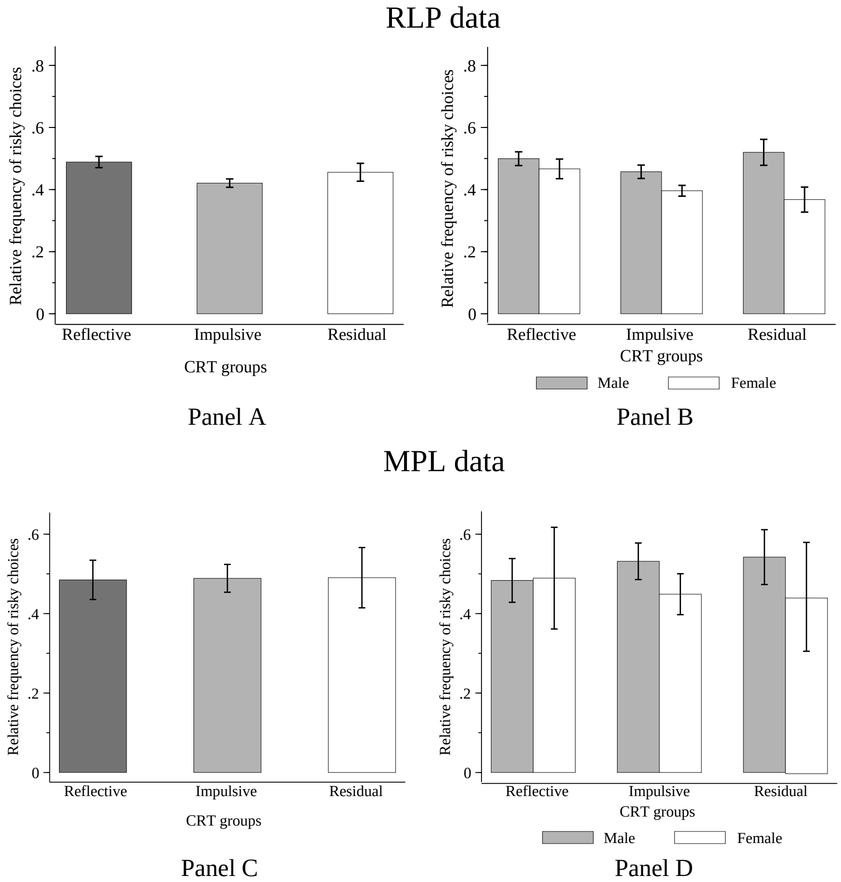


Figure 4: Relative frequency of risky choices in RLP and in MPL data by CRT group, with 95% confidence intervals. Panel A and C (B and D): full sample (by gender).

Panel A in Figure 4 displays the relative frequency of "risky" choices in RLP, where the latter are identified by the higher-variance lottery within the pair. Panel B shows the same information disaggregated by gender. These results confirm, by and large, the commonplace in the literature, that is, that higher cognitive reflection is associated with lower risk aversion (Donkers et al., 2001; Frederick, 2005; Benjamin et al., 2013). More precisely, Panel A shows that *reflective are less risk averse than impulsive*, while the difference between reflective and the residual group seems less important. In addition, Panel B in Figure 4 shows that, once we split our subject pool by gender, females tend to be more risk averse than males within the same CRT group. Besides, for men there are no significant differences in risk aversion across groups, while for women it is higher for reflective group than for the others. This evidence suggests that *both* cognitive ability and gender play an important role in explaining subjects' risk attitudes.

Panel C in Figure 4, displays the relative frequency of risky choices in MPL (i.e., the lottery that yields a 50-50 chance to get all or nothing) for those subjects whose behavior satisfies minimal "consistency conditions", that will be explained and discussed in Section 6.2. Panel D shows the same information disaggregated by gender.

As Panel C shows, aggregate behavior of all CRT groups is almost identical. However, when we disaggregate by gender, we see that risk aversion slightly decreases moving from the reflective to the residual group for males, while this pattern is exactly reversed for females. We also observe that the relative frequency of risky choices for reflective subjects is the same for males and females, although females' choices have higher variability.

There is a caveat here. The summary statistics of Figure 4 neglects relevant features of the underlying economic decisions at stake. When selecting a lottery, subjects most likely compare the profitability of each decision, not simply its associated risk. Put differently, the relative frequency of risky choices does not characterize precisely the economic trade-off underlying both the RLP and the MPL decisions. For this reason, we test the robustness of the preliminary evidence of Figure 4 by estimating, by maximum likelihood, subjects' individual Constant Relative Risk Aversion parameter, ρ, where subjects' choices are assumed to maximize the expected value of the utility function $u(x)$ over monetary prizes x in equation (1), where higher ρ is associated with higher risk aversion (Andersen et al., 2008).

$$u(x) = \frac{x^{1-\rho}}{1-\rho}, \rho \neq 1, \tag{1}$$

Table 4 reports the estimated coefficients using RLP (MPL) data on the left (right) panel, respectively. As for the RLP data (left panel) the estimated coefficients are always are greater than zero and highly significant, which shows that risk aversion is the representative preference for all CRT groups. When we test for the differences in risk aversion across CRT groups, the p-values at the bottom of the table show that it is only significant between reflective and impulsive subjects. When we test for the differences across CRT groups by gender, we find that the overall difference between reflective and impulsive subjects is mainly driven by females. We also find a significant difference between reflective and residual females that is hidden in the aggregate estimations.[16]

As for the MPL data (right panel), the p-values at the bottom of Table 4 show that, at the aggregate level (first column), differences in risk aversion across CRT groups are not significant, thus confirming the preliminary evidence in Figure 4. The same result also holds when we disaggregate by gender, suggesting that the trends we observe in Figure 4 are not statistically significant.

	Random lottery pairs (RLP) protocol			Multiple price list (MPL) protocol		
	All	Males	Females	All	Males	Females
Reflective (R)	0.508***	0.481***	0.545***	0.217***	0.198***	0.223**
	(0.023)	(0.029)	(0.035)	(0.054)	(0.063)	(0.113)
Impulsive (I)	0.571***	0.506***	0.609***	0.188***	0.068	0.296***
	(0.015)	(0.031)	(0.016)	(0.045)	(0.064)	(0.058)
Residual (RS)	0.502***	0.394***	0.627***	0.179**	0.103	0.264**
	(0.047)	(0.080)	(0.031)	(0.078)	(0.081)	(0.128)
P-val R = I	0.012**	0.512	0.081*	0.643	0.117	0.538
P-val R = RS	0.914	0.297	0.065*	0.667	0.323	0.801
P-val I = RS	0.154	0.179	0.592	0.914	0.709	0.806
Obs.s	4,608	2,184	2,424	3,969	2,184	1,785

Table 4: Structural estimation of risk aversion (ρ) using data from RLP and MPL protocols.
Maximum likelihood estimates. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. P-values are from t-statistics to test the hypothesis that the difference in risk aversion between two CRT groups is equal to zero.

5 CRT and Social Preferences

The relationship between cognitive ability and social preferences is, to some extent, yet to be explored. Chen et al. (2013) find that subjects who perform better in the Math portion of the SAT are more generous in both the Dictator game and in a series of small-stakes dictatorial decisions known as *Social Value Orientation* (SVO). In contrast, Ben-Ner et al. (2004) find that the performance in the *Wonderlic test* is weakly and negatively correlated with giving, especially for females.[17] Benjamin

[16]The sign and significance of risk aversion estimated parameters and their differences by CRT groups in Table 4 are unchanged if the female dummy is added as independent variable, as shown in Table C1 in the Appendix. The same holds if estimates are obtained by using linear regressions, as shown in Table C2.

[17]The Wonderlic test score is positively correlated with various measures of intelligence (Hawkins et al., 1990). See footnote 1 for its definition.

et al. (2013) find, instead, that school test scores are not correlated with giving. Somewhat related, Hauge et al. (2009) study the relationship between attitudes to giving in different pro-social tasks (e.g., charitable giving, Dictator Games, etc...) and "cognitive load", which they measure by asking subjects to memorize numbers of 7 digits, some of which are easy (hard) to remember, e.g., 1111111 or 1234567 (9325867 or 7591802). They find that the correlation between cognitive load and giving is small.

Our distributional data are drawn from Project 4 and consist of a sequence of decisions over four monetary payoff pairs in which the identity of the best-paid player is constant across choices. Since choices are not naturally ordered, we provide some descriptive evidence of this experimental environment by introducing an ad hoc index, borrowed from Project 6, which measures the share of the pie allocated to the Dictator (conditional on the specific round choice set):

$$EgoIndex(k) = \frac{x_D(k) - min(x_D(h))}{max_h(x_D(h)) - min_h(x_D(h))}, \tag{2}$$

where $x_D(k)$ denotes the monetary payoff allocated to the Dictator according to option k. In other words, if the Dictator gives him/herself the maximum (minimum) prize available (regardless of what the Recipient obtains), the value of the $EgoIndex(.)$ is 1 (0), respectively.

Figure 5 reports descriptive statistics of the distribution of *EgoIndex*, disaggregated by CRT group and gender. It shows that impulsive (especially female) subjects have higher distributional concerns, with no noticeable difference between reflective and residual subjects. However, we cannot exclude that differences in distributional concerns by CRT group are driven by differences in subjects' ability, in the light of the positive correlation between CRT performance and achievement in ability or school tests observed in the literature.

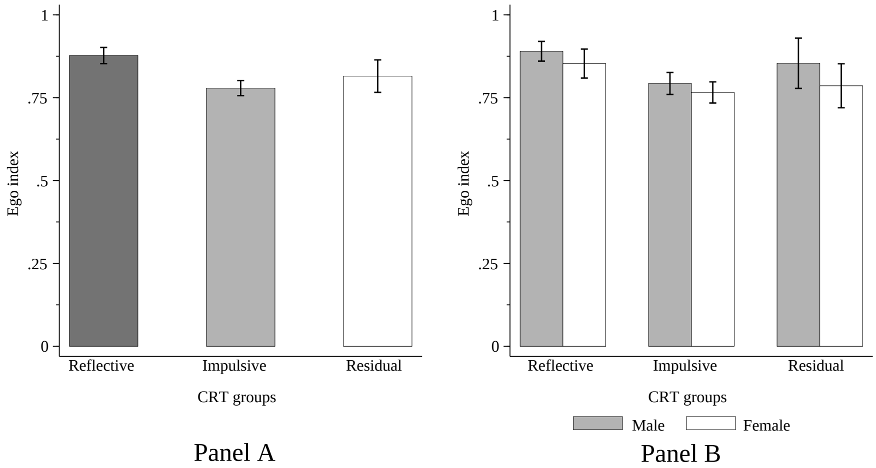


Figure 5: *EgoIndex* by CRT group, with 95% confidence intervals. Panel A (B): full sample (by gender).

Before assessing the empirical content of this preliminary evidence, notice that, by analogy with what we have just discussed for risky choices, Figure 5 captures the economic trade-off underlying Dictators' decisions only partially, as it is calculated looking at the Dictator's payoffs only, and not at the Recipient's. This contrasts with the common view which models social preferences by measuring *relative comparisons* between the Dictator's and the Recipient's payoffs.

For this reason, we test the robustness of the preliminary evidence of Figure Figure 5 by estimating, by maximum likelihood, the classic Fehr and Schmidt (1999) model of social preferences, according to which the Dictator's utility associated with option γ, $u(\gamma)$, not only depends on her own monetary payoff, $x_D(\gamma)$, but also on that of the Recipient, $x_R(\gamma)$, as follows:

$$u(\gamma) = x_D(\gamma) - \alpha max\left\{x_R(\gamma) - x_D(\gamma), 0\right\} - \beta max\left\{x_D(\gamma) - x_R(\gamma), 0\right\}, \quad (3)$$

where the values of α and β determine the Dictator's *envy* (i.e. aversion to inequality when receiving less than the Recipient) and *guilt* (i.e., aversion to inequality when receiving more than the Recipient), respectively.[18] We estimate α and β by using a multinomial logit model in which the utility associated with the Dictator's choice of allocation, γ, follows equation (3). We obtain the estimates by maximum

[18] Our data format seems ideal to identify envy and guilt, in that the identity of the best (worst) paid agent is constant across options.

likelihood and by clustering standard errors at the subject level.

	All		Male		Female	
	α	β	α	β	α	β
Reflective (R)	0.116**	0.533***	0.125**	0.521***	0.0995	0.578***
	(0.048)	(0.047)	(0.053)	(0.051)	(0.083)	(0.096)
Impulsive (I)	0.295***	0.760***	0.272***	0.728***	0.331***	0.789***
	(0.036)	(0.037)	(0.047)	(0.062)	(0.049)	(0.045)
Residual (RS)	0.237***	0.582***	0.130*	0.415***	0.307***	0.665***
	(0.071)	(0.087)	(0.078)	(0.101)	(0.107)	(0.123)
Obs.	8,064	8,064	4,152	4,152	3,912	3,912
P-val. R = I	0.003***	0.000***	0.042***	0.012***	0.016***	0.052*
P-val. R = RS	0.176	0.626	0.955	0.369	0.162	0.582
P-val. I = RS	0.441	0.068*	0.124	0.009***	0.835	0.357

Table 5: Social preferences by CRT group: Fehr and Schmidt (1999)'s structural estimation.
Maximum likelihood estimates. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. P-values are from t-statistics to test the hypothesis that the difference in risk aversion between two CRT groups is equal to zero.

The estimates in Table 5 are all positive and significant across the three CRT groups, indicating inequity aversion (i.e., positive envy and guilt) as the predominant behavior (Cabrales et al., 2010). Our estimates also show highly significant correlation between our CRT partition dummies and the model's estimated coefficients: when we test pairwise differences in the estimates between CRT groups (see bottom of Table 5), we find that impulsive subjects have higher distributional concerns than reflective ones, ($p = 0.003$). In addition, we find that they are also weakly more guilty than the residual group ($p = 0.068$), and this is mostly driven by males' behavior ($p = 0.009$).[19]

The observed differences in social preferences by CRT group, particularly between reflective and impulsive subjects, can be rationalised by the prediction that dual cognitive systems drive individuals' decisions (Kahneman, 2011). This is supported by evidence that subjects with high CRT score are less inclined to behavioural biases than those with low CRT (Bergman et al., 2010; Hoppe and Kusterer, 2011; Oechssler et al., 2009) and by related evidence that subjects' altruism is correlated with their 2D:4D (Brañas-Garza et al., 2013).

In a companion paper, Ponti and Rodriguez-Lara (2014) use data from Project

[19]The sign and significance of social preferences estimated parameters and their differences by CRT groups in Table 5 are unchanged if the female dummy is added as independent variable, as shown in Table C3 in the Appendix.

2 on a Linear Dictator Game of 98 subjects and condition the estimates of Fehr and Schmidt (1999) model one the same CRT group partition used in this paper. They also find that *inequality aversion is typical of impulsive subjects* in "standard" Dictator Games (where Dictators' and Recipients' payoffs are negatively related). By contrast, reflective subjects are associated with negligible social concerns, with the exception of *a higher unconditional altruistic attitude*, i.e., negative envy and positive guilt, in situations where the Dictator's payoff is held constant.

6 Is CRT another rationality test ?

In this section we study whether CRT scores and groups are related with measures of "consistency" associated with subjects' behavior in the experiments, as well as alternative proxies of subjects' cognitive ability. As for the former, our indicators of consistency are related with the lottery choices in RLP and MPL experiments. As for the latter, we consider two additional measures of cognitive ability: *educational achievement* and *financial literacy*. Even though these two measures may depend on many factors, we follow Frederick (2005)'s intuition that certain aspects of cognitive ability, such as reading comprehension and mathematical skills, may aid performance in CRT and are likely to correlate with educational achievement and financial literacy, too.

6.1 Consistency in lottery choices

In this section we test whether cognitive reflection is related with subjects' *consistency* across lottery choices by using both MPL and RLP data. As for MPL, a "consistent" subject is defined as one whose choices satisfy these conditions:

1. She should always choose Lottery B (A) in Decision 1 (21) in the sequence. This condition is due to first-order stochastic dominance.

2. She should switch from Option B to Option A only once in the sequence. This is due to monotonicity and transitivity.

This joint condition partitions our subject pool into two subgroups of (in)consistent subjects, respectively. In a similar vein, another proxy for consistency can be derived by counting the number of switches observed for any given individual, with "inconsistency" growing with the number of switches.

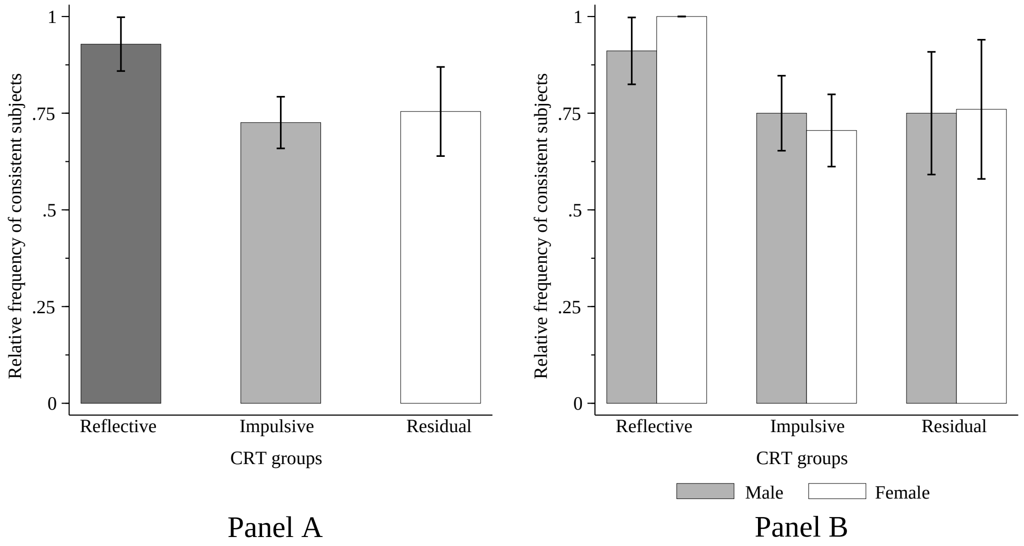


Figure 6: Consistent subjects in lottery choices by CRT group, with 95% confidence intervals.

Figure 6 shows the relative frequency of consistent subjects by CRT group for the full sample (Panel A) and by gender (Panel B). As Figure 6 shows, about 90% of reflective subjects are consistent. This frequency falls to 75% for the other two groups. The 95% confidence intervals in Figure 6 show that reflective subjects are significantly more consistent than any of the other groups, while the difference between the other two subgroups is not significant. Also notice that we do not observe significant gender differences in consistency within each CRT group.

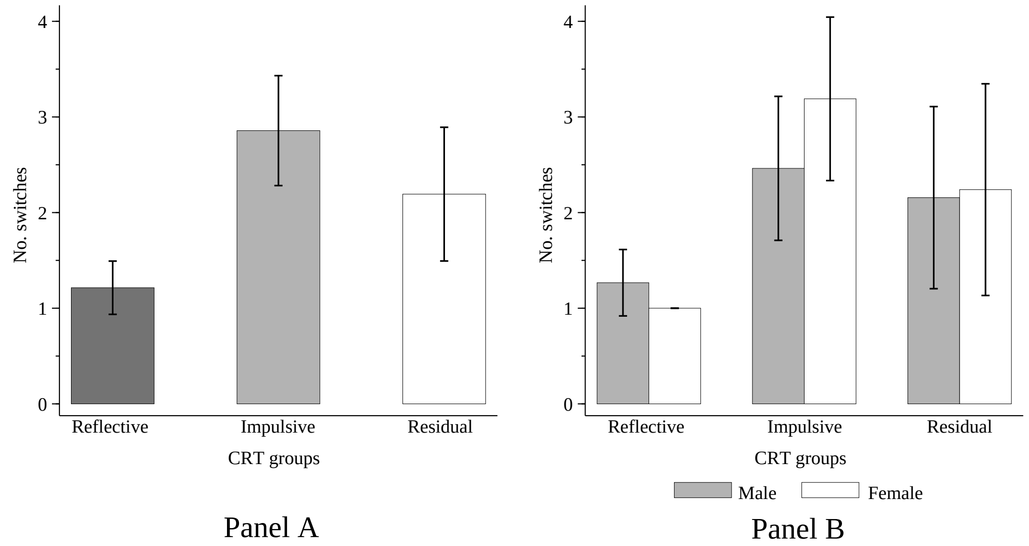


Figure 7: Number of switches in lottery choices by CRT group and gender, with 95% confidence intervals.

Figure 7 shows distribution of mean switches for the full sample (Panel A) and by gender (Panel B). By analogy with Figure 6, the number of switches for the reflective group is significantly smaller than those of the the other two groups, for which we do not detect a significant difference. Again, we do not detect significant gender differences within each CRT group.[20]

Also RLP data provide a relatively straightforward consistency test, in that there are two decisions (out of 24) in which lotteries can be ranked by first-order stochastic dominance. In this respect, "consistent" subjects should never go for the dominated lottery, independently on their degree of risk aversion, ρ, and -actually- for a much broader family of behavioral models of choice under risk than expected utility maximization. Looking at our RLP data, we found that no reflective subject (out of 33) is inconsistent according to our definition, while we found 4 (out of 128, 3%) within the impulsive group and other 4 (out of 31, 13%) within the residual group. To Mann-Whitney standards, these differences are significant, except that between reflective and implulsive.[21]

To summarize, our data clearly show that reflective (residual) subjects are more (less) likely to act consistently in our lottery tasks, respectively, with no detectable

[20] Table C4 in the Appendix reports Mann-Whitney-Wilcoxon tests for pairwise comparisons across CRT groups, both for the full sample and by gender. Table C5 reports the same tests for gender differences. Results are in line with those reported here.

[21] Incidentally, among the 8 inconsistent subjects, there are 5 males and 3 females.

gender effect.

6.2 Grades and financial literacy

Extensive evidence documents that educational achievement is positively correlated with labor market outcomes (Heckman et al., 2006). Similarly, financial literacy has been shown to correlate with stockholding (Christelis et al., 2010) and is an increasingly important objective in high school curricula (Mandell and Klein, 2009).

Dependent variable: number of correct answers in the CRT

	(1)	(2)	(3)	(4)
GPA	0.019*	0.022**	0.010	0.009
	(0.010)	(0.010)	(0.023)	(0.023)
Female		-1.141***		-1.447***
		(0.200)		(0.473)
Financial Literacy			0.573**	0.312
			(0.240)	(0.265)
Observations	432	432	96	96

Table 6: CRT, GPA and financial literacy.
Ordered Logit estimates. Robust standard errors in parentheses. * p<0.1; ** p<0.05; *** p<0.01

We borrow Lusardi and Mitchell (2014) test of financial literacy, which consists of 3 questions on subjects' general knowledge of financial markets. Consistently with Frederick (2005), the ordered logit estimates in Table 6 show that the GPA coefficient -which we measure using subjects' grades at university from 0 to 100- is statistically significant. Also financial literacy is also positively and significantly correlated with CRT. However, after controlling for gender the effect is no longer significant. It seems that the aggregate correlation between CRT and financial literacy is driven by the fact that females in our sample have lower financial literacy.[22]

7 Discussion

Overall, our results confirm *a strong gender component in CRT performance.* With regards to other individual characteristics, we find significant, although quantitatively much smaller, correlations between CRT and 2D:4D, personality traits and family education.

[22] After performing Mann-Whitney test for gender differences, we find that financial literacy is significantly lower for females ($z = 3.588$, p-value $= 0.0003$)

We have also studied whether cognitive reflection is correlated with risk and social preferences. Our structural estimations with RLP data show that reflective subjects tend to be less risk averse than impulsive ones, especially for females. By contrast, MPL data show no significant difference by CRT group or gender, in line with the criticism of Andersson et al. (2013).[23] As for social preferences, impulsive subjects are more envious and guilty than reflective ones, and impulsive males are more guilty than the residual group, while females are not. This evidence complements the findings in Ponti and Rodriguez-Lara (2014) who employ the Dictator Game data of Project 2 and find that, once again, impulsive subjects are those whose behavior markedly differs from that of the other two groups (again, in the direction of inequity aversion).

Finally, we have studied the correlation between cognitive reflection and alternative proxies of cognitive ability. Here we have found that reflective subjects are more likely to satisfy basic consistency requirements in their lottery choice, in contrast with the other two groups (especially, the residual), which are, instead, more prone to violate such conditions. In line with Frederick (2005), we have also found that academic performance (GPA) is positively correlated with CRT. Similar considerations hold for financial literacy, which is also correlated with CRT. However, in this case, the effect seems to be uniquely driven by the underlying gender difference. Additional experimental sessions seem required to increase the low sample size and obtain more robust evidence with respect to this result.

As mentioned in the introduction, it is worth emphasizing that CRT provides not only a measure of cognitive ability, but also of impulsiveness. Furthermore, our analysis shows that it is significantly correlated with various individual characteristics, as well as with alternative measures of cognitive ability and literacy. This leaves the interpretation of our results regarding CRT and economic behavior somewhat open. Of course, one possibility would be to incorporate further explanatory variables in the analysis, allowing us to examine which factors captured by CRT turn out to explain individual heterogeneity in behavior. For example, it would be interesting to check whether the correlation between risk aversion and CRT holds

[23] See also Charness et al. (2013); Filippin and Crosetto (2014) for a discussion of the relative advantages and disadvantages of different risk elicitation protocols.

after controlling for financial literacy, or whether its association with social preferences remains after the inclusion of personality traits and alternative measures of rationality. Unfortunately, the structure of our data is such that we do not have enough observations to perform these types of tests.

By the same token, the observed gender difference in CRT scores remains open to interpretation. The existing literature agrees on a strong gender difference in CRT but, to the best of our knowledge, does not provide an explanation for this finding. Our own evidence as well as that of earlier studies (e.g. Frederick, 2005; Bosch-Domènech et al., 2014) suggest that this difference remains after controlling for a number of individual characteristics such as personality, education, prenatal exposure to testosterone, mathematical ability, etc. One important factor that has received limited attention in the literature regards the incentive structure under which the test is administered, that is, whether or not subjects are rewarded for each correct answer. This could be important if females turn out to have less intrinsic motivation to perform well in this test.[24]

We have only found a few studies that compare CRT performance by gender, checking whether the test is incentivized or not. Oechssler et al. (2009) look at CRT with incentives and find an average score of 2.2 for males and 1.7 for females. Hoppe and Kusterer (2011) also look at CRT with incentives and find scores of 2.12 and 1.61 for males and females, respectively. On the other hand Bosch-Domènech et al. (2014) look at CRT without incentives and find average scores of 0.95 and 0.58 for males and females, respectively. These latter figures are much closer to ours (1.08 and 0.55 for males and females, respectively) than the rest of the cited references, which suggests that *gender differences in performance may be reduced when the CRT is incentivized*.

We conclude by recalling that this paper exploits the richness of our dataset only partially, with particular reference to our behavioral data, in that it focuses on individual decision tasks (mainly related with risk and social preferences).

The link between cognitive reflection and behavior in strategic environments is being studied elsewhere (take, for example, projects 1, 2, 4, 6 or 7). For instance,

[24]One possible reason for females' lower intrinsic motivation may be that they perceive the CRT as a male task.

Ponti and Carbone (2009) find a negative correlation between CRT scores and the level of noise of subjects' play in an experimental model of informational cascades, while Ponti et al. (2014b), within the setting of a simple principal-agent model with moral hazard, show that reflective principals offer higher wages, which, in turn, yield higher effort levels and profits. By the same token, reflective agents exert more effort, which also results in higher expected profits in the experiment.[25]

Moving to a rather different behavioral domain, Ferrara et al. (2015) find that sleep deprivation makes reflective subjects more likely to choose riskier lotteries and induce a more altruistic behavior. By contrast, Albano et al. (2014) do not detect significant differences across CRT groups in both winning probabilities or expected profits in an experimental procurement auction. A more detailed study to relate such a dispersed evidence is currently under way.

[25]More evidence on the interaction between CRT performance and strategic behavior can be found in other articles of this special issue. Benito-Ostolaza et al. (2015), for example, find that high scoring subjects in the Raven's test play more strategically in coordination games. Jones et al. (2015) find that high-CRT people tend to reciprocate more in the second round of the classical Prisoner's Dilemma. Baghestanian and Frey (2014) find that high-CRT GO players tend to be more cooperative in a series of classical games. Lohse (2015) finds that high-CRT people contribute more in a classical one-shot public good game. Interestingly, this effect disappears when they have little time to make their decisions.

References

Albano, G., Di Paolo, R., Ponti, G., and Sparro, M. (2014). Absolute vs. Relative Scoring in Experimental Procurement. mimeo, LUISS Guido Carli Roma.

Alos-Ferrer, C. and Hügelschäfer, S. (2015). Faith in Intuition and Cognitive Reflection. *Journal of Behavioral and Experimental Economics THIS ISSUE.*

Andersen, S., Harrison, G. W., Lau, M. I., and Rutström, E. E. (2008). Eliciting Risk and Time Preferences. *Econometrica*, 76(3):583–618.

Andersson, O., Tyran, J. R., Wengström, E., and Holm, H. J. (2013). Risk Aversion Relates to Cognitive Ability: Fact or Fiction? Working Papers 2013:9, Lund University, Department of Economics.

Apicella, C. L., Dreber, A., Campbell, B., Gray, P. B., Hoffman, M., and Little, A. C. (2008). Testosterone and Financial Risk Preferences. *Evolution and Human Behavior*, 29(6):384–390.

Baghestanian, S. and Frey, S. (2014). GO Figure: Analytic and Strategic Skills are Separable. *Journal of Behavioral and Experimental Economics THIS ISSUE.*

Barrick, M. R. and Mount, M. K. (1991). The Big Five Personality Dimensions and Job Performance: A Meta-analysis. *Personnel Psychology*, 44(1):1–26.

Ben-Ner, A., Kong, F., and Putterman, L. (2004). Share and Share Alike? Gender-pairing, Personality, and Cognitive Ability as Determinants of Giving. *Journal of Economic Psychology*, 25(5):581–589.

Benet-Martinez, V. and John, O. P. (1998). Los Cinco Grandes Across Cultures and Ethnic Groups: Multitrait-multimethod Analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*, 75(3):729–750.

Benito-Ostolaza, J. M., Hernández, P., and Sanchis-Llopis, J. A. (2015). Are Individuals with Higher Cognitive Ability Expected to Play More Strategically? *Journal of Behavioral and Experimental Economics THIS ISSUE.*

Benjamin, D. J., Brown, S. A., and Shapiro, J. M. (2013). Who is "Behavioral"? Cognitive Ability and Anomalous Preferences. *Journal of the European Economic Association*, 11(6):1231–1255.

Bergman, O., Ellingsen, T., Johannesson, M., and Svensson, C. (2010). Anchoring and Cognitive Ability. *Economics Letters*, 107(1):66–68.

Borghans, L., Duckworth, A. L., Heckman, J. J., and Ter Weel, B. (2008a). The Economics and Psychology of Personality Traits. *Journal of Human Resources*, 43(4):972–1059.

Borghans, L., Golsteyn, B. H. H., Heckman, J. J., and Meijers, H. (2009). Gender Differences in Risk Aversion and Ambiguity Aversion. *Journal of the European Economic Association*, 7(2-3):649–658.

Borghans, L., Meijers, H., and Ter Weel, B. (2008b). The Role Of Noncognitive Skills In Explaining Cognitive Test Scores. *Economic Inquiry*, 46(1):2–12.

Bosch-Domènech, A., Brañas-Garza, P., and Espín, A. M. (2014). Can Exposure to Prenatal Sex Hormones (2D: 4D) Predict Cognitive Reflection? *Psychoneuroendocrinology*, 43:1–10.

Brañas-Garza, P., García-Muñoz, T., and González, R. H. (2012). Cognitive Effort in the Beauty Contest Game. *Journal of Economic Behavior & Organization*, 83(2):254–260.

Brañas-Garza, P., Kovářík, J., and Neyse, L. (2013). Second-to-Fourth Digit Ratio Has a Non-Monotonic Impact on Altruism. *PLoS ONE*, 8(4):e60419.

Brañas-Garza, P. and Rustichini, A. (2011). Organizing Effects of Testosterone and Economic Behavior: Not Just Risk Taking. *PloS ONE*, 6(12):e29842.

Cabrales, A., Miniaci, R., Piovesan, M., and Ponti, G. (2010). Social Preferences and Strategic Uncertainty: An Experiment on Markets and Contracts. *American Economic Review*, 100(5):2261–78.

Charness, G., Gneezy, U., and Imas, A. (2013). Experimental Methods: Eliciting Risk Preferences. *Journal of Economic Behavior & Organization*, 87:43–51.

Chen, C. C., Chiu, I. M., Smith, J., and Yamada, T. (2013). Too Smart to Be Selfish? Measures of Cognitive Ability, Social Preferences, and Consistency. *Journal of Economic Behavior & Organization*, 90(0):112–122.

Christelis, D., Jappelli, T., and Padula, M. (2010). Cognitive Abilities and Portfolio Choice. *European Economic Review*, 54(1):18–38.

Coates, J. M., Gurnell, M., and Rustichini, A. (2009). Second-to-fourth Digit Ratio Predicts Success Among High-frequency Financial Traders. *Proceedings of the National Academy of Sciences*, 106(2):623–628.

Cueva, C., Iturbe-Ormaetxe, I., Ponti, G., and Tomás, J. (2014). The Disposition Effect with Transaction Costs: An Experimental Analysis. mimeo, Universidad de Alicante.

Daly, M., Harmon, C. P., and Delaney, L. (2009). Psychological and Biological Foundations of Time Preference. *Journal of the European Economic Association*, 7(2-3):659–669.

Del Pozo, X., Galliera, A., Ponti, G., and Sikora, I. (2013). Social Preferences, Risk Preferences and the Hexagon Condition. mimeo, Universidad de Alicante.

Di Cagno, D., Harrison, G. W., Miniaci, R., and Ponti, G. (2014). Social Preferences over Utilities. mimeo, LUISS Guido Carli Roma.

Digman, J. M. (1990). Personality Structure: Emergence of the Five-factor Model. *Annual Review of Psychology*, 41(1):417–440.

Dohmen, T., Falk, A., Huffman, D., and Sunde, U. (2010). Are Risk Aversion and Impatience Related to Cognitive Ability? *American Economic Review*, 100(3):1238–1260.

Donkers, B., Melenberg, B., and Van Soest, A. (2001). Estimating Risk Attitudes Using Lotteries: A Large Sample Approach. *Journal of Risk and Uncertainty*, 22(2):165–195.

Evans, J. S. B. T. (1984). Heuristic and Analytic Processes in Reasoning. *British Journal of Psychology*, 75(4):451–468.

Fehr, E. and Schmidt, K. M. (1999). A Theory of Fairness, Competition and Cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Ferrara, M., Bottasso, A., Tempesta, D., Carrieri, M., De Gennaro, L., and Ponti, G. (2015). Gender Differences in Sleep Deprivation Effects on Risk and Inequality Aversion: Evidence from an Economic Experiment. *PLoS ONE*, 10(3):e0120029.

Filippin, A. and Crosetto, P. (2014). A Reconsideration of Gender Differences in Risk Attitudes. IZA Discussion Papers 8184, Institute for the Study of Labor (IZA).

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

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4):25–42.

Gill, D. and Prowse, V. (2015). Cognitive Ability, Character Skills, and Learning to Play Equilibrium: A Level-k Analysis. *Journal of Political Economy forthcoming.*

Grimm, V. and Mengel, F. (2012). An Experiment on Learning in a Multiple Games Environment. *Journal of Economic Theory*, 147(6):2220–2259.

Harrison, G. W. and McDaniel, T. (2008). Voting Games and Computational Complexity. *Oxford Economic Papers*, 60(3):546–565.

Hauge, K. E., Brekke, K. A., Johansson, L., Johansson-Stenman, O., and Svedsäter, H. (2009). Are Social Preferences Skin Deep? Dictators under Cognitive Load. Working Papers in Economics 371, University of Gothenburg, Department of Economics.

Hawkins, K. A., Faraone, S. V., Pepple, J. R., Seidman, L. J., and Tsuang, M. T. (1990). WAIS-R Validation of the Wonderlic Personnel Test as a Brief Intelligence Measure in a Psychiatric Sample. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, 2(2):198–201.

Heckman, J. J. and LaFontaine, P. A. (2010). The American High School Graduation Rate: Trends and Levels. *The Review of Economics and Statistics*, 92(2):244–262.

Heckman, J. J. and Rubinstein, Y. (2001). The Importance of Noncognitive Skills: Lessons from the GED Testing Program. *American Economic Review*, 91(2):145–149.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics*, 24(3):411–482.

Hey, J. D. and Orme, C. (1994). Investigating Generalizations of Expected Utility Theory Using Experimental Data. *Econometrica*, 62(6):1291–1326.

Hoppe, E. I. and Kusterer, D. J. (2011). Behavioral Biases and Cognitive Reflection. *Economics Letters*, 110(2):97 – 100.

Insler, M., Compton, J., and Schmitt, P. (2015). The Investment Decisions of Young Adults Under Relaxed Borrowing Constraints. *Journal of Behavioral and Experimental Economics THIS ISSUE.*

Jang, K. L., Livesley, W. J., and Vemon, P. A. (1996). Heritability of the Big Five Personality Dimensions and Their Facets: A Twin Study. *Journal of Personality*, 64(3):577–592.

John, O. P., Naumann, L. P., and Soto, C. J. (2008). Paradigm Shift to the Integrative Big Five Trait Taxonomy. In John, O. P., Robins, R. W., and Pervin, L. A., editors, *Handbook of Personality: Theory and Research*, volume 3, pages 114–158. Guilford Press New York, NY, third edition.

John, O. P. and Srivastava, S. (1999). The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. In Pervin, L. A. and John, O. P., editors, *Handbook of Personality: Theory and Research*, pages 102–138. Guilford Press New York, NY, second edition.

Jones, G., al Ubaydli, O., and Jaap Weel, B. (2015). Average Player Traits as Predictors of Cooperation in a Repeated Prisoner's Dilemma. *Journal of Behavioral and Experimental Economics THIS ISSUE.*

Judge, T. A., Higgins, C. A., Thoresen, C. J., and Barrick, M. R. (1999). The Big Five Personality Traits, General Mental Ability, and Career Success Across the Life Span. *Personnel Psychology*, 52(3):621–652.

Kahneman, D. (2011). *Thinking, Fast and Slow.* Macmillan.

Kirby, K. N., Winston, G. C., and Santiesteban, M. (2005). Impatience and Grades: Delay-discount Rates Correlate Negatively with College GPA. *Learning and Individual Differences*, 15(3):213–222.

Kiss, H. J., Rodriguez-Lara, I., and Rosa-García, A. (2015). Think Twice Before Running! Bank Runs and Cognitive Abilities. *Journal of Behavioral and Experimental Economics THIS ISSUE*.

Kruskal, W. H. and Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Loehlin, J. C. (1992). *Genes and Environment in Personality Development*. Sage Publications, Inc.

Loehlin, J. C., McCrae, R. R., Costa Jr, P. T., and John, O. P. (1998). Heritabilities of Common and Measure-specific Components of the Big Five Personality Factors. *Journal of Research in Personality*, 32(4):431–453.

Lohse, J. (2015). Smart or Selfish - When Smart Guys Finish Nice. *Journal of Behavioral and Experimental Economics THIS ISSUE*.

Lusardi, A. and Mitchell, O. S. (2014). The Economic Importance of Financial Literacy: Theory and Evidence. *Journal of Economic Literature*, 52(1):5–44.

Mandell, L. and Klein, L. S. (2009). The Impact of Financial Literacy Education on Subsequent Financial Behavior. *Journal of Financial Counseling and Planning*, 20(1):15–24.

Manning, J. T., Scutt, D., Wilson, J., and Lewis-Jones, D. I. (1998). The Ratio of 2nd to 4th Digit Length: A Predictor of Sperm numbers and Concentrations of Testosterone, Luteinizing Hormone and Oestrogen. *Human Reproduction*, 13(11):3000–3004.

Morsanyi, K., Busdraghi, C., and Primi, C. (2014). Mathematical Anxiety is Linked to Reduced Cognitive Reflection: A Potential Road from Discomfort in the Mathematics Classroom to Susceptibility to Biases. *Behavioral and Brain Functions*, 10(1):31.

Neyse, L. and Brañas-Garza, P. (2014). Digit Ratio Measurement Guide. MPRA Paper 54134, University Library of Munich, Germany.

Noussair, C., Tucker, S., and Xu, Y. (2014). A Future Market Reduces Bubbles but Allows Greater Profit for More Sophisticated Traders. Discussion Paper 2014-051, Tilburg University, Center for Economic Research.

Oechssler, J., Roider, A., and Schmitz, P. W. (2009). Cognitive Abilities and Behavioral Biases. *Journal of Economic Behavior & Organization*, 72(1):147–152.

Pearson, M. and Schipper, B. C. (2012). The Visible Hand: Finger Ratio (2D: 4D) and Competitive Bidding. *Experimental Economics*, 15(3):510–529.

Ponti, G. and Carbone, E. (2009). Positional Learning with Noise. *Research in Economics*, 63(4):225–241.

Ponti, G. and Rodriguez-Lara, I. (2014). Social Preferences and Cognitive Reflection: Evidence from Dictator Game Experiment. mimeo, LUISS GUIDO Carli Roma.

Ponti, G., Rodriguez-Lara, I., and Di Cagno, D. (2014a). Doing it Now or Later with Payoff externalities: Experiental Evidence on Social Time Preference. mimeo, LUISS GUIDO Carli Roma.

Ponti, G., Sartarelli, M., Sykora, I., and Zhukova, V. (2014b). The Price of Entrepreneurship. Evidence from the Lab. mimeo, Universidad de Alicante.

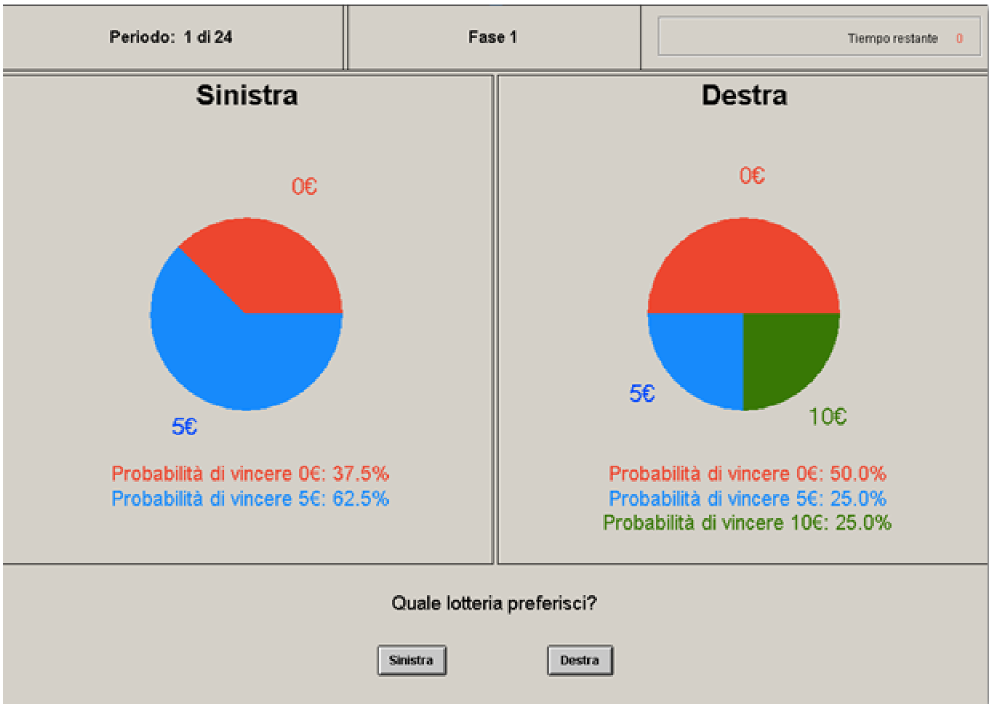
Sapienza, P., Zingales, L., and Maestripieri, D. (2009). Gender Differences in Financial Risk Aversion and Career Choices Are Affected by Testosterone. *Proceedings of the National Academy of Sciences*, 106(36):15268–15273.

Toplak, M. E., West, R. F., and Stanovich, K. E. (2011). The Cognitive Reflection Test as a Predictor of Performance on Heuristics-and-biases Tasks. *Memory & Cognition*, 39(7):1275–1289.

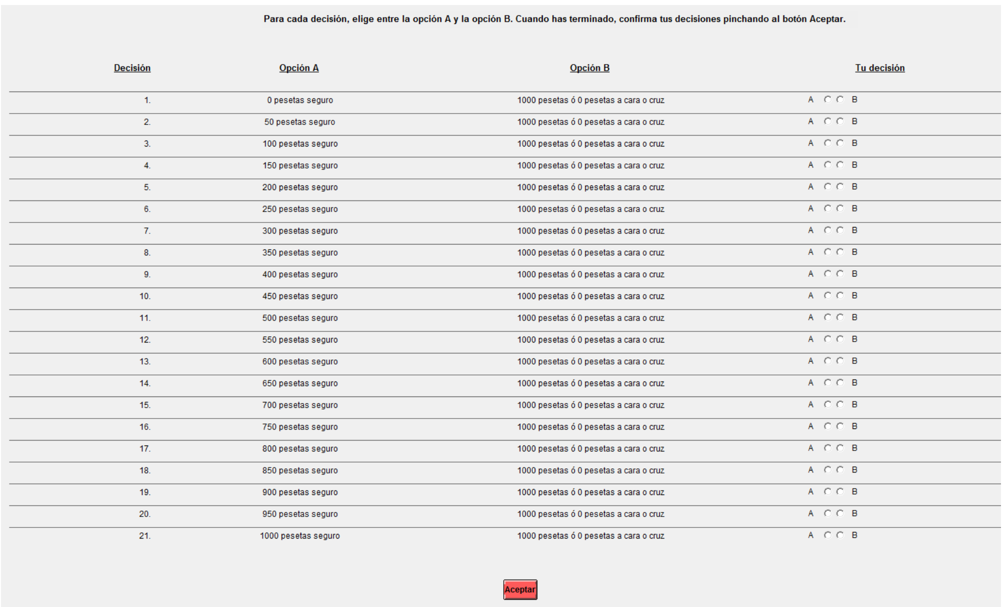
Zhao, H. and Seibert, S. E. (2006). The Big Five Personality Dimensions and Entrepreneurial Status: A Meta-analytical Review. *Journal of Applied Psychology*, 91(2):259–271.

Appendix (not for publication)

Appendix A



Panel A



Panel B

Figure A1: Panel A: user interface of the RLP (Project 3). Panel B: user interface of the MPL (Projects 4 and 8).

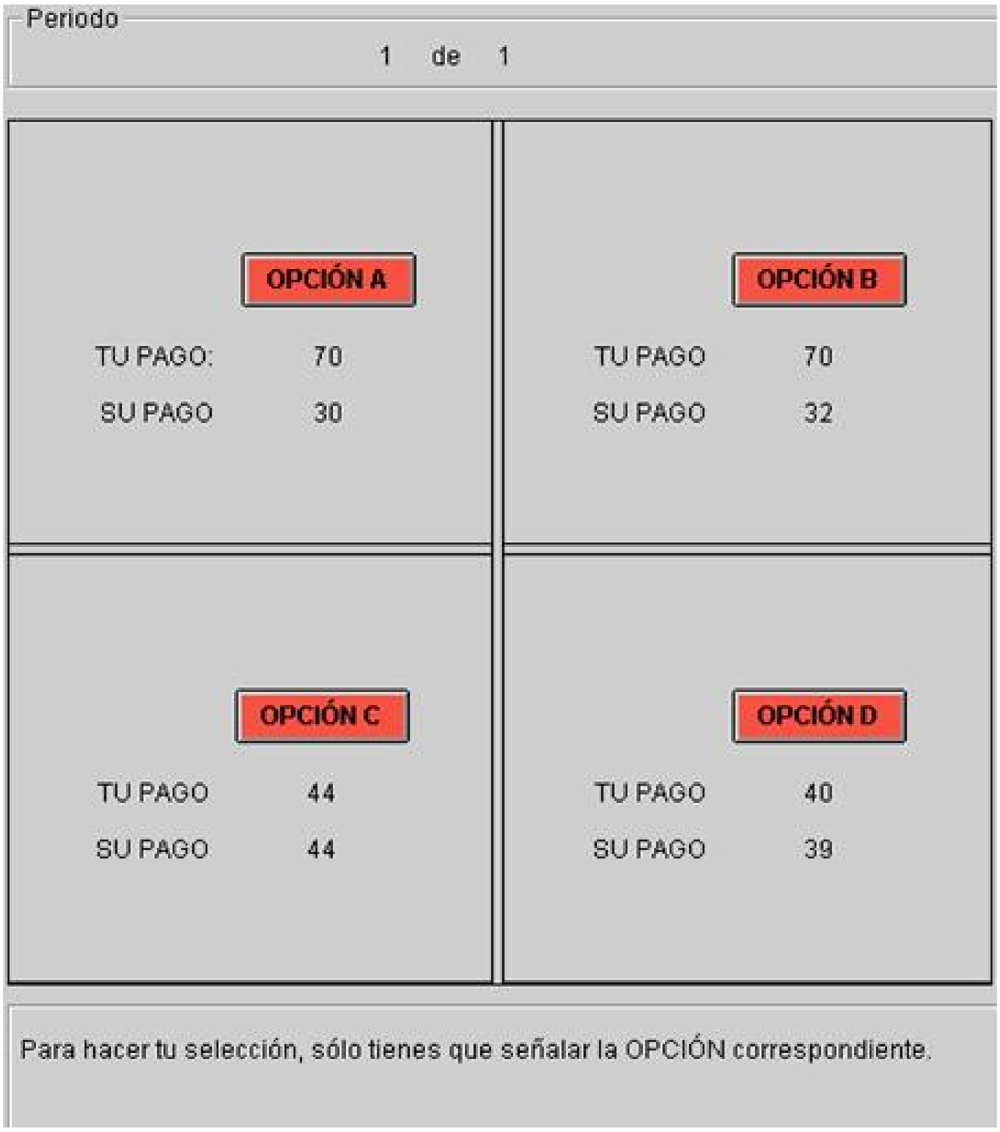


Figure A2: Distributional task, user interface.

Appendix B

Personality trait	Definition
Openness	Being open to new ideas and intellectually curious, imaginative, nonconforming, unconventional an autonomous
Neuroticism	Tendency to experience psychological distress, exhibit poor emotional adjustment and experience negative affects, such as anxiety, insecurity and hostility
Agreeableness	Tendency to be compassionate, cooperative, trusting, compliant, caring and gently
Conscientiousness	Tendency to show control and self-discipline, is comprised on two related facets: achievement and dependability
Extraversion	Pronounced engagement with outside world, it represents the tendency to be sociable, assertive, active and experience positive affects such as energy and zeal

Table B1: Big 5 personality traits

	Female				Male			
	Mean			Kruskal-Wallis	Mean			Kruskal-Wallis
	Reflective	Impulsive	Residual	p-value	Reflective	Impulsive	Residual	p-value
Left hand 2D:4D	0.981	0.986	0.993	0.668	0.965	0.976	0.970	0.050**
Rightt hand 2D:4D	0.975	0.984	0.989	0.494	0.964	0.972	0.971	0.366
Neuroticism	0.538	0.548	0.506	0.612	0.394	0.459	0.418	0.035**
Extraversion	0.601	0.576	0.617	0.497	0.574	0.645	0.553	0.000***
Openness	0.773	0.682	0.686	0.007***	0.706	0.714	0.677	0.008***
Agreeableness	0.727	0.679	0.685	0.324	0.681	0.692	0.650	0.001***
Conscientiousness	0.731	0.688	0.702	0.382	0.672	0.688	0.661	0.148
Family education	0.446	0.394	0.473	0.295	0.521	0.370	0.488	0.002***
(1+ parent uni. degree)								
N. languages >2	0.414	0.430	0.471	0.885	0.453	0.297	0.286	0.080*

Table B2: Means of individuals' characteristics and p-values of Kruskal-Wallis test of differences among CRT groups. *** p <0.01, ** p<0.05, * p<0.1.

	Full sample			Female			Male		
	Reflective -		Impulsive -	Reflective -		Impulsive -	Reflective -		Impulsive -
	Impulsive	Residual	Residual	Impulsive	Residual	Residual	Impulsive	Residual	Residual
Left hand 2D:4D	0.011***	0.014**	0.366	0.830	0.415	0.417	0.022**	0.080*	0.649
Right hand 2D:4D	0.025**	0.073*	0.792	0.339	0.208	0.653	0.144	0.575	0.843
Neuroticism	0.002***	0.069*	0.485	0.997	0.381	0.352	0.015**	0.051*	0.893
Extraversion	0.321	0.574	0.071*	0.486	0.877	0.287	0.025**	0.210	0.000***
Openness	0.070*	0.005***	0.031**	0.002***	0.022**	0.917	0.808	0.014**	0.002***
Agreeableness	0.573	0.023**	0.009***	0.134	0.291	0.789	0.721	0.004***	0.000***
Conscientiousness	0.981	0.252	0.271	0.160	0.434	0.730	0.413	0.187	0.062*
Family education (1+ parent uni. degree)	0.001***	0.815	0.014**	0.342	0.701	0.160	0.001***	0.607	0.058*
N. languages >2	0.214	0.508	0.781	0.876	0.654	0.664	0.033**	0.134	0.907

Table B3: Mann-Whitney-Wilcoxon p-values of differences in means of individuals' characteristics among CRT groups. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

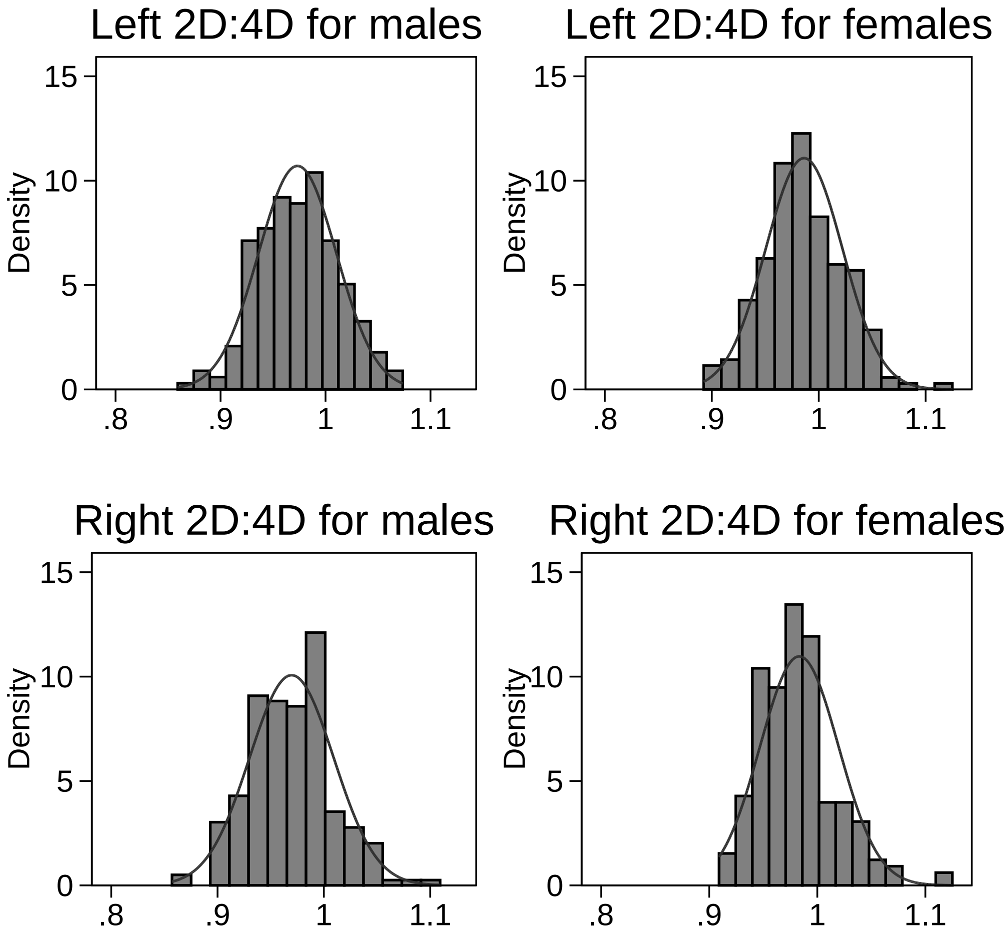


Figure B1: Second to fourth digit ratio (2D:4D) histogram by gender.

Appendix C

	Random lottery pairs (RLP) protocol			Multiple price list (MPL) protocol		
Female	0.606***		0.090***	0.236***		0.099*
	(0.014)		(0.025)	(0.051)		(0.056)
Reflective (R)		0.508***	0.468***		0.217***	0.183***
		(0.023)	(0.028)		(0.054)	(0.060)
Impulsive (I)		0.571***	0.511***		0.188***	0.141***
		(0.015)	(0.026)		(0.045)	(0.048)
Residual (RS)		0.502***	0.494***		0.179**	0.131
		(0.047)	(0.034)		(0.078)	(0.083)
P-val R = I		0.012**	0.093*		0.643	0.520
P-val R = RS		0.914	0.483		0.667	0.583
P-val I = RS		0.154	0.618		0.914	0.910
Obs.s	9,168	9,216	9,168	3,969	3,969	3,969

Table C1: Risk aversion by CRT group: structural estimation using data from RLP and MPL protocols.

Maximum likelihood estimates. Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. P-values are from t-statistics to test the hypothesis that the difference in risk aversion between two CRT groups is equal to zero. The total number of observations is the product between the number of subjects and the number of lottery choices per subject.

	Random lottery pairs (RLP) protocol				Multiple price list (MPL) protocol			
Female	-0.075***		-0.064***	-0.061**	-0.065**		-0.069**	-0.083**
	(0.018)		(0.019)	(0.025)	(0.027)		(0.028)	(0.035)
Reflective (R)		0.068***	0.051**	0.042		-0.001	-0.020	-0.044
		(0.020)	(0.021)	(0.028)		(0.030)	(0.031)	(0.035)
Residual (RS)		0.035	0.017	0.063		0.002	0.002	0.011
		(0.030)	(0.029)	(0.044)		(0.041)	(0.039)	(0.039)
Female * R				0.028				0.085
				(0.042)				(0.071)
Female * RG				-0.091				-0.017
				(0.057)				(0.078)
Constant	0.483***	0.421***	0.459***	0.457***	0.518***	0.489***	0.525***	0.532***
	(0.013)	(0.012)	(0.017)	(0.021)	(0.016)	(0.018)	(0.021)	(0.023)
P-val R = RS		0.304	0.276	0.637		0.955	0.615	0.181
P-val R = I fem.				0.024**				0.511
P-val R = RS fem.			0.276	0.023**				0.574
Obs.s	382	384	382	382	186	186	186	186

Table C2: OLS estimation of risky choices using data from RLP and MPL protocols. Means of risky choices by subjects over the number of lotteries played are used. In RLP subjects play 24 lotteries and the risky option in each of them is the one with the highest variance. In MPL the subjects play 21 lotteries and the share of risky choices is computed as the relative frequency of risky lottery options chosen. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. P-values are from t-statistics to test the hypothesis that the difference in risk aversion between two CRT groups is equal to zero. The total number of observations is the product between the number of subjects and the number of lottery choices per subject.

	(1)	(2)	(3)
	Coefficients for α		
Female	0.295***		0.059
	(0.041)		(0.048)
Reflective (R)		0.116**	0.092*
		(0.048)	(0.049)
Impulsive (I)		0.295***	0.267***
		(0.036)	(0.046)
Residual (RS)		0.237***	0.200***
		(0.071)	(0.078)
P-val R=I		0.003***	0.005***
P-val R=RS		0.177	0.234
P-val I=RS		0.441	0.379
	Coefficients for β		
Female	0.724***		0.115**
	(0.038)		(0.058)
Reflective (R)		0.533***	0.494***
		(0.047)	(0.054)
Impulsive (I)		0.760***	0.700***
		(0.037)	(0.054)
Residual (RS)		0.582***	0.514***
		(0.087)	(0.089)
P-val R=I		0.000***	0.000***
P-val R=RS		0.626	0.841
P-val I=RS		0.068*	0.054*
Obs.	8,064	8,064	8,064

Table C3: Social preferences by CRT group: Fehr and Schmidt (1999)'s structural estimation.

Maximum likelihood estimates. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. P-values are from t-statistics to test the hypothesis that the difference in social preferences between two CRT groups is equal to zero.

	Relative frequency of consistent subjects				Number of switches		
	Reflective	Impulsive	Residual		Reflective	Impulsive	Residual
Reflective	.	.	.	Reflective	.	.	.
Impulsive	0.002***	.	.	Impulsive	0.005**	.	.
Residual	0.013***	0.672	.	Residual	0.108	0.326	.

(a) Full sample

	Reflective	Impulsive	Residual		Reflective	Impulsive	Residual
Reflective	.	.	.	Reflective	.	.	.
Impulsive	0.031**	.	.	Impulsive	0.033**	.	.
Residual	0.060*	1.000	.	Residual	0.230	0.608	.

(b) Male

	Reflective	Impulsive	Residual		Reflective	Impulsive	Residual
Reflective	.	.	.	Reflective	.	.	.
Impulsive	0.039**	.	.	Impulsive	0.113	.	.
Residual	0.088*	0.591	.	Residual	0.306	0.408	.

(c) Female

Table C4: P-values of Mann-Whitney-Wilcoxon tests of relative frequency of consistent subjects and number of switches for pairs of CRT groups. *p-value<0.1, **p-value<0.05, ***p-value<0.01

	Relative frequency of consistent subjects	Number of switches
Reflective	0.314	0.613
Impulsive	0.511	0.511
Residual group	0.932	0.947

Table C5: P-values of Mann-Whitney-Wilcoxon tests of gender differences in the relative frequency of consistent subjects and number of switches by CRT group. *p-value<0.1, **p-value<0.05, ***p-value<0.01