Journal of Behavioral and Experimental Economics 000 (2015) 1–13

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/socec

Cognitive (ir)reflection: New experimental evidence☆

Carlos Cueva[a], Iñigo Iturbe-Ormaetxe[a], Esther Mata-Pérez[a], Giovanni Ponti[a,b,*], Marcello Sartarelli[a], Haihan Yu[a], Vita Zhukova[a]

[a] Universidad de Alicante, Spain
[b] LUISS Guido Carli Roma, Dipartimento di Economia e Finanza, LUISS Guido Carli, Viale Romania 32, 00185 Roma, Italy

ARTICLE INFO

Article history:
Received 29 November 2014
Revised 2 September 2015
Accepted 4 September 2015
Available online xxx

JEL Classification:
C91
D81
J16

Keywords:
Behavioral economics
Cognitive reflection
Gender effects
Experiments

ABSTRACT

We study how cognitive abilities correlate with behavioral choices by collecting evidence from almost 1200 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.

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, Winston, and Santiesteban, 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, Meijers, and Ter Weel, 2008b) to the Wonderlic test, aimed at assessing problem-solving ability (Ben-Ner, Kong, and Putterman, 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 min to make 5 widgets, how long would it take 100 machines to make 100 widgets? (*Correct answer*: 5 min).

☆ 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.

* Corresponding author at: LUISS Guido Carli Roma, Dipartimento di Economia e Finanza, LUISS Guido Carli, Viale Romania 32, 00185 Roma, Italy. Tel.: +39 06 85225740.
E-mail address: giuba@ua.es, gponti@luiss.it (G. Ponti).

[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).

http://dx.doi.org/10.1016/j.socec.2015.09.002
2214-8043/

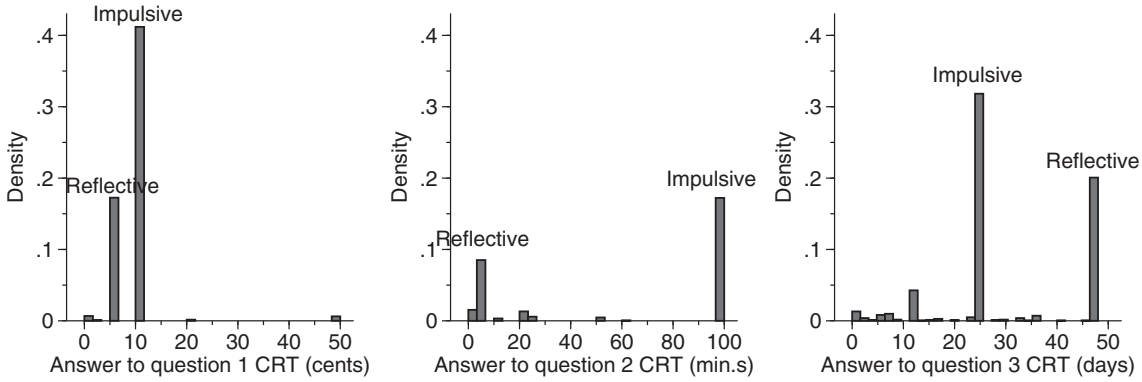

Fig. 1. CRT answers distributions.

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 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, Brañas-Garza, and Espín, 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, Roider, and Schmitz, 2009; Alos-Ferrer and Hügelschäfer, 2015; Noussair, Tucker, and Xu, 2014; Kiss, Rodriguez-Lara, and Rosa-García, 2015; Insler, Compton, and Schmitt, 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, Roider, and Schmitz, 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, Busdraghi, and Primi, 2014; Toplak, West, and Stanovich, 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 1200 observations (see Section 2 for a detailed description). To get directly into the discussion around which this paper is built, Fig. 1 reports the distribution of CRT answers of our compound dataset. As Fig. 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).

Fig. 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, García-Muñoz, and González, 2012; Grimm and Mengel, 2012).

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

$$iCRT = 1(\text{CRT1} = 10) + 1(\text{CRT2} = 100) + 1(\text{CRT3} = 24),$$

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 Fig. 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 Fig. 2, while the first two groups have a strong gender component, the latter distributes across genders almost equally.

Please cite this article as: C. Cueva et al., Cognitive (ir)reflection: New experimental evidence, Journal of Behavioral and Experimental Economics (2015), http://dx.doi.org/10.1016/j.socec.2015.09.002