2015

The Cognitive Basis of Social Behavior: Cognitive Reflection Overrides Antisocial but Not Always Prosocial Motives

Brice Corgnet
Chapman University, corgnet@chapman.edu

Antonio M. Espín
Middlesex University Business School

Roberto Hérnan-Gonzalez
Universidad de Granada

The Cognitive Basis of Social Behavior: Cognitive Reflection Overrides Antisocial but Not Always Prosocial Motives

The cognitive basis of social behavior: cognitive reflection overrides antisocial but not always prosocial motives

Brice Corgnet[1], Antonio M. Espín[2,3], Roberto Hernán-González[3,4]

Author affiliations:

1. Economic Science Institute, Argyros School of Business and Economics, Chapman University, CA 92866, Orange, USA.

2. Economics Department, Middlesex University Business School, London NW4 4BT, UK.

3. Granada Lab of Behavioral Economics (GLoBE), Universidad de Granada, Granada 18071, Spain.

4. Business School, University of Nottingham, Nottingham NG8 1BB, UK

Abstract

Even though human social behavior has received considerable scientific attention in the last decades, its cognitive underpinnings are still poorly understood. Applying a dual-process framework to the study of social preferences, we show in two studies that individuals with a more reflective/deliberative cognitive style, as measured by scores on the Cognitive Reflection Test (CRT), are more likely to make choices consistent with "mild" altruism in simple non-strategic decisions. Such choices increase social welfare by increasing the other person's payoff at very low or no cost for the individual. The choices of less reflective individuals (i.e. those who rely more heavily on intuition), on the other hand, are more likely to be associated with either egalitarian or spiteful motives. We also identify a negative link between reflection and choices characterized by "strong" altruism, but this result holds only in Study 2. Moreover, we provide evidence that the relationship between social preferences and CRT scores is not driven by general intelligence. We discuss how our results can reconcile some previous conflicting findings on the cognitive basis of social behavior.

Keywords: dual-process; reflection; intuition; social preferences; altruism; spitefulness; prosocial behavior; antisocial behavior; inequality aversion

Introduction

Mounting evidence shows that humans cooperate with non-kin even when doing so implies paying irrecoverable costs (Ledyard 1995, Gintis 2000, Henrich et al. 2001, Fehr & Gächter 2002, Bowles & Gintis 2003, Camerer 2003). These prosocial behaviors are inconsistent with the strict pursue of self-interest and thus constitute a challenge for disciplines ranging from evolutionary biology to the social and behavioral sciences (Nowak 2006, Fehr & Camerer 2007,

Harbaugh et al. 2007, Perc & Szolnoki 2008, Roca et al. 2009, Capraro 2013, Exadaktylos et al. 2013, Rand & Nowak 2013, Gutiérrez-Roig et al. 2014, Raihani 2014).

In recent years, the cognitive underpinnings of social behavior have been increasingly studied, as their understanding is key for building a comprehensive account of the proximate—and, indirectly, also ultimate—explanations of human sociality (Stevens & Hauser 2004, Rand & Nowak 2013, Zaki & Mitchell 2013). Much of the advances on this front have been made within the framework of dual-process theories, which point to the existence of an interaction between fast, automatic/intuitive ("System 1") and slow, controlled/reflective ("System 2") decision making processes (Hogarth 2001, Stanovich 2010, Kahneman 2011). From this perspective, most research has focused on answering the question of whether human prosocial (as opposed to selfish) behavior is the result of intuition or reflection (Loewenstein & O'Donoghue 2004, Moore & Loewenstein 2004, Rand et al. 2012, Zaki & Mitchell 2013). In other words, are humans' automatic responses selfish or prosocial?

An extensive research program on the topic has identified cooperation as the intuitive response in anonymous one-shot social dilemma experiments, with further reflection leading to more selfish choices (Rand et al. 2012, 2014a, 2014b, Cone & Rand 2014, Evans et al. 2014, Rand & Kraft-Todd 2014). These findings have led to the Social Heuristics Hypothesis (SHH; Rand et al. 2014b), according to which people internalize social behaviors that generate personal benefits in daily life. In contrast to most economic experiments, daily life interactions are often repeated and face-to-face, and this implies that behaving cooperatively may be rewarding in the long run (through reciprocity, reputation or due to the existence of sanctions; Hamilton 1964, Williams 1966, Trivers 1971, Fudenberg & Maskin 1986, Bowles & Gintis 2003). Individuals interacting in environments where helping others usually pays off would thus be more likely to internalize prosocial behaviors than individuals dwelling more "inhospitable" environments (Rand et al. 2012, Peysakhovich & Rand 2015). Such internalization would lead people to apply prosocial heuristics even in situations where cooperation is maladaptive, such as in one-shot anonymous economic experiments.

Even though the SHH has received considerable empirical support (Roch et al. 2000, Cornelissen et al. 2011, Rand et al. 2012, 2014a, 2014b, Lotito et al. 2013, Cone & Rand 2014, Evans et al. 2014, Nielsen et al. 2014, Rand & Kraft-Todd 2014, Schultz et al. 2014), a number of findings seem inconsistent with the idea of spontaneous prosociality and calculated selfishness (e.g. Knoch et al. 2006, 2010, DeWall et al. 2008, Piovesan & Wengström 2009, Martinsson et al. 2012, 2014, Tinghog et al. 2013, Crockett et al. 2014, Jaber-López et al. 2014, Verkoeijen & Bouwmeester 2014).[1] In this paper, we shall argue that our understanding of the sources of these apparent contradictions may benefit from an in-depth analysis of the motivations underlying social behavior. A distinction should thus be made between observed behavioral outcomes and underlying social motivations (Falk et al. 2005, Jensen 2010, Espín et al. 2012, Brañas-Garza et al. 2014,). Indeed, a variety of "prosocial" motivations (e.g. altruism or egalitarianism; see below) can trigger seemingly identical prosocial behaviors. It might be the case that some of the prosocial motivations that account for a specific behavior are linked to intuition whereas others are linked to reflection. This may explain why the analysis of isolated social decisions has led to mixed findings regarding the role of intuitive and reflective processes in prosocial behavior. Our main argument is that identifying the precise driving forces underlying a given social behavior is a first and necessary step toward understanding its cognitive underpinnings.

The previous discussion focused on the often-studied prosocial side of human behavior but it nonetheless extends to the less-studied antisocial side. Evidence from economic experiments also shows that people often make "antisocial" decisions that reduce others' welfare without any apparent personal gain (Zizzo & Oswald 2001, Fehr & Gächter 2002, Knoch et al. 2006, Herrman & Orzen 2008, Herrmann et al. 2008, Abbink et al. 2010, Espín et al. 2012, Kimbrough & Reiss 2012, Brañas-Garza et al. 2014). Spiteful behaviors that harm others even at one's own cost may yet be advantageous, for example, in social environments where survival hinges upon one's relative standing in the group.[2] Therefore, following the SHH argument, some people might internalize behaviors that not only promote but also reduce others' welfare as an adaptation to their daily life interactions. Welfare-reducing behaviors are likely to respond to

[1] See the General Discussion section for a more detailed overview of some of these findings.
[2] This may be the case, e.g., when there exist high levels of environmental uncertainty/volatility (e.g. violence) or when local (vs. global) competition for resources prevails (Gardner & West 2004, McCullough et al. 2013, Sylwester et al. 2013, Prediger et al. 2014).

antisocial motives that aim at increasing one's relative standing (Kirchsteiger 1994, Van Lange 1999, Charness & Rabin 2002, Jensen 2012). This logic has been applied, for instance, to understanding the punishment decisions of non-cooperators in social dilemma games (Shinada et al. 2004, Falk et al. 2005, Gächter & Herrmann 2011, Espín et al. 2012). When the punishing individual is a cooperator, however, fairness-based explanations are often put forward (Fehr & Schmidt 1999, Fehr & Gächter 2002, Gächter & Herrmann 2009, Espín et al. 2012). From this viewpoint, fairness concerns, which are traditionally considered to be prosocial (Van Lange 1999), can also lead to behaviors that reduce the payoff of another individual.

To analyze the cognitive underpinnings of human social interaction, it is thus important to distinguish people's actual behaviors and motivations. To do so, it is necessary to bring back the too-often ignored antisocial motivations at the center of the debate.

Disentangling social motives

Research on social (other-regarding) preferences has developed theoretical models aimed at explaining, at the proximate level, non-selfish behaviors observed in one-shot economic experiments, such as cooperation in social dilemmas or the rejection of "unfair" offers in bargaining games (Rabin 1993, Kirchsteiger 1994, Levine 1998, Fehr & Schmidt 1999, Bolton & Ockenfels 2000, Charness & Rabin 2002, Falk & Fischbacher 2006, Cox et al. 2007, López-Pérez 2008). Social preferences models can be grouped into three broad classes (Fehr & Schmidt 2006): intentions-based (e.g. Rabin 1993), type-based (e.g. Levine 1998) and outcome-based (e.g. Fehr & Schmidt 1999) preferences. In this paper, we will focus on "outcome-based"—or distributional—social preferences which introduce the payoffs of relevant others into the individuals' utility functions. In other words, individuals with outcome-based social preferences behave *as if* they were maximizing a utility function which includes a concern for the payoff of others, in addition to their own payoff. Models of social preferences constitute a potent tool to systematically characterize the motivations behind individuals' decisions in social environments.[3]

[3] Note that we are using preference types as a classification device, irrespective of whether or not these are completely stable characteristics of individuals as often assumed in the social preferences literature. Indeed a number of studies challenge such interpretation and suggest that social preferences can be modulated/manipulated (e.g. Rand et al. 2012, McCall et al. 2014).

From a theoretical standpoint, we will base our analyses on the inequality-aversion model of Fehr & Schmidt (1999), which accounts for a potential asymmetry between advantageous and disadvantageous payoff comparisons between the self and a referent other (e.g. Loewenstein et al. 1989). We extend the previous model so as to capture behaviors than may not strictly follow from the standard inequality-aversion model. We will rely on a generalized and flexible specification of preferences that will allow us to disentangle competing explanations of individuals' decisions, including both prosocial and antisocial motivations. Similar approaches have been followed for instance by Charness & Rabin (2002), Engelmann & Strobel (2004), Engelmann (2012) and Cox (2013).

According to the basic specification of the model (Fehr & Schmidt 1999), the utility derived by individual *i* from the payoff vector $X=(x_1, \ldots, x_n)$ is given by:

$$U_i(X) = x_i - \alpha_i \frac{1}{n-1}\sum_{j \neq i} \max\{x_j - x_i, 0\} - \beta_i \frac{1}{n-1}\sum_{j \neq i} \max\{x_i - x_j, 0\}, \quad (1)$$

where the parameters α_i and β_i refer to the individual *i*'s *aversion* to disadvantageous (i.e. "envy") and advantageous inequality (i.e. "compassion"), respectively. Thus, a self-regarding individual who is indifferent to others' payoffs would exhibit $\alpha_i = \beta_i = 0$. A person with other-regarding motives would prefer either to increase or decrease others' payoffs depending on the sign and value of α_i [β_i] if others' payoffs are above [below] her own payoffs.

One caveat in the categorization of social behavior is that individuals' decisions in standard economic games are typically consistent with different types of motivations. For instance, both spiteful and selfish motives would identically lead to zero transfers in dictator games (Brañas-Garza et al. 2014). Similarly, the acceptance of a low offer in the ultimatum game could result from either selfishness or altruism (Staffiero et al. 2013). In order to uncover the driving forces behind a particular decision, a clear cut procedure is to observe the decisions made by the same individual in different social situations (Falk et al. 2005, Espín et al. 2012, Yamagishi et al. 2012, Staffiero et al. 2013, Brañas-Garza et al. 2014, Peysakhovich et al. 2014). In addition, these decisions should be free of strategic or reciprocal concerns since these could alter behavior and distort the assessment of outcome-based preferences (Charness & Rabin 2002). To infer the social motives underlying choices we will study a series of two-player allocation decisions for

which an individual decides how to allocate different amounts of money between herself and a passive recipient. In our two-person case, the model presented in Eq. (1) is reduced to:

$$U_i(x_i, x_j) = x_i - \alpha_i \max\{x_j - x_i, 0\} - \beta_i \max\{x_i - x_j, 0\}. \quad (2)$$

In its original formulation, Fehr & Schmidt (1999) assume $\alpha_i \geq \beta_i \geq 0$, which means that individuals can be either egalitarian ($\alpha_i \geq 0$ and $\beta_i \geq 0$; with at least one inequality being strict) or selfish ($\alpha_i = \beta_i = 0$). This parameterization also implies that people are assumed to display at least as envy as compassion ($\alpha_i \geq \beta_i$). We do not impose these restrictions on the model parameters so that individuals' motivations can be characterized as follows:

(i) *Self-interest* if individuals' decisions maximize their own payoff ($\alpha_i = 0$ and $\beta_i = 0$);

(ii) *Altruism* if individuals' decisions maximize the other's payoff ($\alpha_i \leq 0$ and $\beta_i \geq 0$; with at least one inequality being strict)—a concern for *social welfare* also applies if, in addition, $|\alpha_i|, |\beta_i| < 0.5$ (Engelmann 2012)[4]—;

(iii) *Egalitarianism* if individuals' decisions minimize payoff inequality ($\alpha_i \geq 0$ and $\beta_i \geq 0$; with at least one inequality being strict);

(iv) *Spitefulness* if individuals' decisions minimize the other's payoff ($\alpha_i \geq 0$ and $\beta_i \leq 0$; with at least one inequality being strict)—which, for empirically relevant values of α_i and β_i, also implies a preference for increasing the individual's relative standing.

(v) *Inequality-seeking* if individuals' decisions maximize payoff inequality ($\alpha_i \leq 0$ and $\beta_i \leq 0$; with at least one inequality being strict)—note that we include this type of preferences for the sake of completeness even though few individuals typically fall into this category.

Hence, we classify individuals' motives according to the combination of both model parameters. Following previous literature, we shall consider that altruism and egalitarianism are prosocial preferences (e.g. Van Lange 1999, Fehr & Schmidt 2006) while spitefulness is antisocial (e.g. Herrmann & Orzen 2008, Jensen 2012, Brañas-Garza et al. 2014).

To illustrate the relevance of the former categorization in understanding the cognitive basis of social behavior, we consider the example of social dilemma experiments which have served as

[4] Note that an individual with $\alpha \leq -0.5$ and $\beta \geq 0.5$ would give money away even when doing so does not increase the total surplus (i.e. social welfare, also referred to as "efficiency").

the basic empirical ground for the SHH. A social dilemma is a situation where the individual and the collective interests are in conflict (e.g. Dawes 1980, Van Lange et al. 2013). In the typical setup, individuals are matched in groups of size *n* and have to decide simultaneously whether to allocate their endowment to a group project (i.e. "cooperate") or whether to keep it for themselves (i.e. "defect"). The resources allocated to the group project are multiplied by a synergy factor *s* and then shared evenly among the *n* group members, regardless of their individual contributions. The social dilemma arises when $1 < s < n$, so that cooperation maximizes the total surplus of the group (i.e. social welfare, the collective interest) whereas defection maximizes individual payoff. However, in addition to maximizing one's own payoff, defection also minimizes others' payoff, so that defection can result either from self-interest or from antisocial, spiteful motives (Falk et al. 2005, Espín et al. 2012). On the other hand, cooperation increases others' payoffs as well as social welfare and can thus result from altruistic motives, either with or without a concern for social welfare. Finally, depending on the level of cooperation one expects from the other group members, both defection and cooperation could also result from egalitarian motives. An individual whose goal is to reduce inequality would defect [cooperate] if she expects others to defect [cooperate]. The previous arguments highlight the difficulty to isolate the social motivations and cognitive underpinnings of people's behavior in social dilemmas. For example, it is difficult to conclude whether reflective defection (e.g. Rand et al. 2012, 2014b) is driven by a desire to (i) maximize one's own payoff (the interpretation advanced by the authors), (ii) reduce others' payoffs or (iii) equalize payoffs with (expectedly) greedy partners. Defection in social dilemmas is thus consistent with selfish, antisocial, and even prosocial motivations. Similarly, it is difficult to disentangle whether intuitive cooperative responses are the result of a willingness to (i) increase others' payoffs, irrespective of the cost, (ii) increase others' payoffs only when the cost is sufficiently low so that social welfare increases or (iii) equalize payoffs with (expectedly) cooperative partners. The number of possible interpretations of people's behavior in social dilemmas would increase even further if we considered reciprocal concerns (i.e. responding kindly to kind actions and unkindly to unkind actions, see e.g. Charness & Rabin 2002). Isolating the motivations of individual behavior presents similar challenges in most of the economic games which are used in the literature on the cognitive basis of social behavior such as standard dictator and ultimatum

games. To alleviate these concerns, our experimental design makes use of several decisions in short, cognitively undemanding and non-strategic tasks.

A trait approach to cognitive reflection

To isolate intuitive and reflective cognitive processes, previous behavioral research on social behavior has primarily relied on the analysis of reaction times (e.g. Rubinstein 2007, Piovesan & Wengström 2009, Brañas-Garza et al. 2012b, Rand et al. 2012, Lotito et al. 2013) and the use of experimental manipulations, such as cognitive load (e.g. Cornelissen et a. 2011, Duffy and Smith 2014, Hauge et al. 2014, Schulz et al. 2014) or time pressure (e.g. Tinghog et al. 2013, Cone & Rand 2014, Rand et al. 2014a, 2014b, Rand & Kraft-Todd 2014). In this paper, we adopt a trait approach which relies on the assumption that individuals who have a more intuitive cognitive style are more likely to make decisions guided by automatic processes (System1), whereas more reflective individuals are more likely driven by deliberative processes (System 2) (Oechssler et al. 2009, Toplak et al. 2011, Peysakhovich & Rand 2015). Subjects' cognitive styles are assessed through the Cognitive Reflection Test (CRT; Frederick 2005), which measures the ability to override intuitive responses and to engage in further reflection before making a decision. The CRT is a short task consisting of a set of insights problems (three in the original form of Frederick 2005, and seven in the extended version introduced by Toplak et al. 2014). The CRT differs from other measures of cognitive abilities as it is designed to prompt an intuitive, yet incorrect, answer to the respondent's mind. To reach the correct answer, the person must override this automatic response by engaging in reflection.

The CRT fits in nicely with the dual-process approach of decision making. The responses to the test are indeed a good proxy for the individuals' tendency to make intuitive vs. reflective decisions. CRT scores have been found to predict one's own ability to refrain from using inaccurate heuristics in a variety of situations (Oechssler et al. 2009, Toplak et al. 2011).[5] Furthermore, there is evidence that the same behaviors that are observed after experimental

[5] Also, the CRT has been found to correlate with one's ability to delay gratification (Frederick, 2005, Bosch-Domènech et al. 2014) and avoid distractions at work (Corgnet et al. 2014b). In addition, student performance in the CRT has been shown to correlate positively with earnings in experimental asset markets (Corgnet et al. 2014a) and other individual tasks involving the capacity to think backwards (Brañas-Garza et al. 2012a). Finally, the CRT has also been shown to correlate negatively with the adoption of paranormal beliefs (Pennycook et al. 2012, Shenhav et al. 2012).

manipulations of intuitive processing covary with CRT scores in the expected direction (e.g. Shenhav et al. 2012). With regards to social behavior, Peysakhovic & Rand (2015) show that an individual's score on the CRT can predict her tendency to apply previously-acquired social heuristics in environments where they are not advantageous. The authors first conducted repeated social dilemmas where cooperation was or was not advantageous before embedding subjects in one-shot games (social dilemma, dictator and trust games) where prosocial behavior was detrimental to subjects' payoff. As predicted by the SHH, subjects who had interacted in the environment where cooperation was advantageous were on average more prosocial in the subsequent one-shot games compared to those who had interacted in the environment were cooperation was disadvantageous. However, after separating subjects according to cognitive style, the authors show that the predicted spillover effect was only observed among subjects with low CRT scores.

Our empirical strategy will be to correlate subjects' answers to the extended version of the CRT (Toplak et al. 2014) with their decisions in a short, cognitively undemanding and non-strategic task aimed at eliciting the social motives underlying behavior. A similar approach has been undertaken in an independent study conducted by Cueva et al. (2015) and Ponti & Rodriguez-Lara (2015). We present the results of two studies one of which was conducted in the US and the other in Spain.

Study 1

Methods

Participants and general protocol. Participants were 150 students (44.67% female; mean age 20.61±2.73 (SD)) from Chapman University in the U.S. These participants were recruited from a database of more than 2,000 students. A subset of the whole database received invitations at random for participating in the current study, which is part of a larger research program on cognitive abilities and economic decision making. The local IRB approved this research. All participants provided informed consent prior to participating. No deception was used.

We conducted a total of 12 sessions, nine of which with 12 participants and three of which with 14 participants. On average, sessions lasted for 45 minutes. All subjects completed the same tasks in the same order. Importantly, since our aim is to study reflection as a cognitive style (i.e. the trait approach), the social preferences elicitation task was performed before the CRT. Otherwise, having completed the CRT could have induced a reflective mindset which might alter the relationship between trait reflectiveness and the behavior under study (Paxton et al. 2012). In any case, in between the social preferences elicitation task and the CRT participants completed a series of unrelated tasks for about 15 minutes and had a break of 10 minutes. This protocol alleviates concerns about the existence of between-tasks spillover effects (e.g. Fromell et al. 2014) which may potentially induce reverse causality.

Cognitive style assessment. We measured the participants' tendency to rely on intuition vs. reflection using the Cognitive Reflection Test introduced by Frederick (2005). To the original CRT questions, we added four questions recently developed by Toplak et al. (2014). The full set of questions can be found in (supplementary) Text S1. In Table S1, we display the % of subjects answering each question correctly, split by gender. As expected, males performed better in the test than females (Frederick 2005, Bosch-Domènech et al. 2014). Our measure of cognitive reflection is given by the total number of correct answers (from 0 to 7). The full distribution of correct answers by males (mean = 3.67±2.25) and females (mean = 2.39±1.95) is provided in Figure S1.

In addition to CRT, we also measured general intelligence which is likely to be a confounding factor of the (potential) relationship between CRT scores and social behavior. Because answering CRT questions require cognitive abilities, CRT scores partly capture general intelligence in addition to cognitive reflection (Frederick 2005, Stanovich 2009). However, cognitive reflection differs from intelligence as measured in standard IQ tests (e.g. Raven matrices). Intelligence tests measure one's capacity to compute solutions to problems but fail to assess one's capacity to engage in reflection (Stanovich 2009). Although basic cognitive abilities are required to answer the CRT correctly, an intelligent person may often rely on automatic answers (System 1) falling short of blocking intuitive processes by engaging in reflection (System 2). In order to evaluate the importance of general intelligence as a possible confound in

the relationship between CRT and social behavior, we measured subjects' IQ using the Raven progressive matrices test (Raven 1941) and used it as a control variable in our analyses. Specifically, we used the odd number of the last three series of matrices (Jaeggi et al. 2010). The number of matrices correctly solved in the Raven test (in our sample, ranging from 8 to 18, mean = 14.61±2.12) is a conventional measure of cognitive ability. This test captures an important aspect of cognitive ability which is referred to as fluid intelligence or algorithmic thinking (Stanovich 2009, 2010).

Consistently with Frederick (2005) and Stanovich (2009, 2010) we find moderate positive correlation between the number of correct answers in the CRT and Raven tests ($r = 0.43$, $p < 0.01$) which suggests that CRT and Raven are not entirely measuring the same cognitive skills. As is standard practice, none of the cognitive tests were incentivized (Frederick 2005).

Social preferences elicitation. We elicited social preferences à la Bartling et al. (2009) by asking participants to make four choices between two possible allocations of money between themselves and another anonymous participant with whom they were randomly matched. All participants made all the four decisions. We used this short task because it provides a good balance between (maximizing) the information that can be obtained and (minimizing) the cognitive effort required to complete the task. In each experimental session, two participants and one of the four decisions were selected at random for payment. The choice of the first participant in the selected decision was used to allocate payoffs between the two participants (e.g. Sheremeta and Shields 2013). All decisions were anonymous.

The allocation decisions are described in Table 1. Option A always yielded an even distribution of money ($2 for both the self and the other participant), whereas option B yielded uneven payoffs. The first two decisions refer to the advantageous domain while the last two decisions refer to the disadvantageous domain. For each decision, we show in parentheses the envy/compassion parameter associated to choosing the egalitarian and non-egalitarian options (i.e. options A and B) and in square brackets the proportion of subjects who chose each option. In order to compute the model parameters, we assume that utility is linear over the range of payoffs involved in the task (Fehr & Schmidt 1999).

Decision #	Option A self, other		Option B self, other	
1	\$2,\$2	($\beta \geq 0$) [86%]	\$2,\$1	($\beta \leq 0$) [14%]
2	\$2,\$2	($\beta \geq 0.5$) [23%]	\$3,\$1	($\beta \leq 0.5$) [77%]
3	\$2,\$2	($\alpha \geq 0$) [42%]	\$2,\$4	($\alpha \leq 0$) [58%]
4	\$2,\$2	($\alpha \geq 0.5$) [31%]	\$3,\$5	($\alpha \leq 0.5$) [69%]

Table 1. Decisions in the social preferences task (Study 1). For each option, we display the payoff for the decision-maker and the recipient, the associated model parameters (in parentheses) and the % of subjects choosing it (in square brackets).

As it happens with nearly every single decision in social interactions, each choice is consistent with multiple social preferences. For instance, in Decision 1 the participants had to decide whether or not to increase the payoff of a worse-off counterpart by \$1 at no cost—or, alternatively, whether or not to reduce the other's payoff below one's own by \$1 at no cost. Choosing option A in Decision 1 implies $\beta \geq 0$ (compassion) and thus it may, depending on the exact value of β and the sign of α, be consistent with either egalitarianism, altruism, social-welfare concerns or self-interest (a selfish individual would choose randomly in this decision). Option B in Decision 1 is associated with $\beta \leq 0$, which means that it can be chosen by individuals driven by either spitefulness or self-interest. Note that Decision 2 resembles the standard dictator game (Forsythe et al. 1994) in the sense that increasing the other's payoff does not increase the total surplus, i.e. social welfare. On the other hand, Decisions 3 and 4 resemble the decision of a second player (responder) in the standard ultimatum game (Güth et al. 1982)—if we leave reciprocal concerns aside—who has to choose whether to reject (option A) or accept (option B) a disadvantageous split proposed by the first player (proposer).

Results and discussion

Decision analysis

Columns (1a)-(4a) of Table 2 display a series of Probit models estimating the likelihood of choosing option B (i.e. the non-egalitarian choice) in each of the four decisions as a function of CRT scores, and controlling for gender. Columns (1b)-(4b) replicate the same regressions but using Raven scores, instead of CRT, as the main explanatory variable. Finally, in columns (1c)-(4c) both CRT and Raven are included as regressors. Robust standard errors clustered at the

individual level are presented in parentheses. In Figure S3, we display the % of subjects choosing option B in each decision, broken down into two CRT groups, namely individuals with below-median (i.e. three or less correct answers, n = 86) and above-median (n = 64) scores.

Dep var:	Decision 1 $\beta \leq 0$ (vs ≥ 0)	Decision 2 $\beta \leq 0.5$ (vs ≥ 0.5)	Decision 3 $\alpha \leq 0$ (vs ≥ 0)	Decision 4 $\alpha \leq 0.5$ (vs ≥ 0.5)
	(1a)	(2a)	(3a)	(4a)
CRT	-0.136**	-0.054	0.249***	0.236***
	(0.065)	(0.054)	(0.055)	(0.059)
female	-0.573*	-0.329	-0.205	-0.219
	(0.306)	(0.236)	(0.224)	(0.233)
cons	-0.465*	1.077***	-0.438*	-0.070
	(0.278)	(0.252)	(0.230)	(0.233)
ll	-57.017	-79.185	-88.038	-81.547
Wald χ^2	5.27*	2.43	23.72***	19.25***
pseudo R^2	0.061	0.014	0.137	0.126
	(1b)	(2b)	(3b)	(4b)
Raven	-0.061	0.034	0.093*	0.121**
	(0.052)	(0.055)	(0.050)	(0.050)
female	-0.438	-0.242	-0.468**	-0.446**
	(0.273)	(0.228)	(0.211)	(0.219)
cons	-0.024	0.373	-0.945	-1.051
	(0.852)	(0.811)	(0.745)	(0.742)
ll	-58.909	-79.473	-97.676	-88.273
Wald χ^2	3.25	1.45	8.33**	9.50***
pseudo R^2	0.030	0.010	0.043	0.054
	(1c)	(2c)	(3c)	(4c)
CRT	-0.134*	-0.090	0.253***	0.221***
	(0.075)	(0.057)	(0.061)	(0.064)
Raven	-0.007	0.074	-0.009	0.036
	(0.070)	(0.059)	(0.055)	(0.055)
female	-0.571*	-0.361	-0.202	-0.233
	(0.309)	(0.234)	(0.226)	(0.236)
cons	-0.376	0.131	-0.319	-0.540
	(0.936)	(0.817)	(0.734)	(0.749)
ll	-57.012	-78.420	-88.026	-81.365
Wald χ^2	5.41	4.57	23.72***	19.30***
pseudo R^2	0.061	0.023	0.137	0.128

Table 2. Non-egalitarian choice (option B) as a function of CRT and Raven (Study 1). Probit estimates. The α and β parameters associated with the dependent variable are displayed on top of each column. In "a" regressions, the main explanatory variable is CRT score. In "b" regressions, the main explanatory variable is Raven score. In "c" regressions, both CRT and Raven scores are included as explanatory variables. Robust standard errors clustered on individuals are shown in parentheses. *, **, *** denote significance at the 0.10, 0.05 and 0.01 level, respectively. N = 150 in all regressions.

From column (1a) of Table 2, we observe that the CRT score is negatively and significantly associated with the choice of option B in Decision 1 ($p = 0.04$), suggesting that more reflective subjects are less likely to reduce the counterpart's payoff below their own payoff. In terms of the model parameters, subjects with higher CRTs are less likely to exhibit $\beta \leq 0$. A two-sided binomial test rejects the hypothesis that above-median CRTs are indifferent between the two options in Decision 1 (i.e. 50% probability of choosing option B, $p < 0.01$), as would be the case for an individual motivated by self-interest (i.e. $\beta = 0$). For below-median CRTs, the binomial test yields a similar result ($p < 0.01$). Therefore, regardless of CRT, most subjects seem to exhibit strictly positive compassion ($\beta > 0$) (see Figure S3).

However, in Decision 2, where increasing the other's payoff is costly, CRT is no longer significant ($p > 0.30$, column 2a). This result suggests that the probability that the compassion parameter exceeds 0.5 does not differ across CRT scores. Additionally, within both the above-median and below-median CRT groups, a two-sided binomial test rejects that subjects are indifferent between the two options (ps < 0.01). This suggests that, regardless of CRT, $\beta \neq 0.5$. Indeed, for both above- and below-median CRTs, the % of subjects choosing option B is strictly above 50%, suggesting a median β strictly below 0.5 (see Figure S3).

Taken together, the results of Decision 1 and 2 indicate that, whereas the majority of subjects exhibit $\beta \in (0, 0.5)$, subjects with lower CRT scores are yet significantly more likely to exhibit $\beta \leq 0$.

With respect to disadvantageous comparisons, column (3a) shows that CRT positively and significantly predicts choosing option B in Decision 3 ($p < 0.01$), which indicates that more reflective individuals are more likely to exhibit $\alpha \leq 0$. From Figure S3, we observe that this effect is strong, as nearly 80% of the subjects with above-median CRT decide not to lower their counterpart's payoff (this is significantly different from 50%: two-sided binomial test, $p < 0.01$), while only about 42% of below-median CRTs do so (which is not significantly different from 50%, $p > 0.16$). These results suggest that high-CRT individuals are not indifferent between both options in Decision 3—as would be the case for an individual motivated by self-interest, i.e. $\alpha =$

0. In sum, high-CRT individuals are mostly characterized by $\alpha < 0$, while the envy parameter that best characterizes low CRT individuals seems to be close to zero or even slightly positive.

The results for Decision 4 are similar to those for Decision 3 as option B is positively and significantly predicted by CRT ($p < 0.01$, column 4a). This suggests that more reflective individuals are also more likely to exhibit $\alpha \leq 0.5$. Observing that more than 80% of the above-median CRT subjects choose option B in Decision 4 (see Figure S3; this proportion is significantly different from 50%: two-sided binomial test, $p < 0.01$), we can conclude that the envy parameter that best describes high CRTs is strictly lower than 0.5. In the case of below-median CRTs, however, this percentage falls to 58% (which is not significantly different from 50%, $p > 0.16$). Following the results of Decisions 3 and 4, low-CRT subjects, on average, display values of α which are apparently higher than those of high-CRT subjects.

Note that the qualitative nature of our statistical results does not depend on whether we use CRT scores or a binary categorization of CRT (as in Figure S3). Using above-median (vs. below-median) CRT as a binary explanatory variable in the regression analysis instead of CRT scores yields similar results (see Table S3). The effect of CRT in Decision 1 is, however, no longer significant at standard levels ($p = 0.21$).

Now, we turn to the second set of regressions of Table 2 (columns 1b-4b), where subjects' choices are estimated as a function of Raven scores. For those decisions for which CRT was found to be a significant predictor (namely Decisions 1, 3 and 4), the effect of Raven is qualitatively similar to that of CRT, although the statistical significance is lower (even non-significant in the case of Decision 1, $p = 0.28$). These results may indicate that a non-negligible share of the observed relationship between CRT and social preferences is actually driven by general intelligence. In order to address this point, we conducted a last series of regressions in which the scores on both cognitive measures are included as explanatory variables (columns 1c-4c). The regression results point to the opposite direction: the effect of CRT remains statistically significant while the significance of Raven scores completely vanishes when both variables are included in the same model. Thus, it is CRT which drives the relationship between Raven and social preferences. As an indication of the strength of this effect, note that the coefficient

associated to Raven scores is reduced by more than 70% after controlling for CRT in the three aforementioned decisions. So, it seems that finding the correct solution in the Raven test involves some level of cognitive reflection, and it is this aspect which entirely explains the relationship between Raven scores and social preferences. Given that CRT accounts for virtually all the effect of Raven on social decisions, from now on, we concentrate on the analysis of CRT scores.

Social preferences categorization

According to the above results, the decisions of most high-CRT individuals can be characterized as non-envious, i.e. $\alpha < 0$, and moderately compassionate, i.e. $\beta \in (0, 0.5)$. Although the majority of low-CRT individuals seem to be moderately compassionate as well they differ from high-CRT individuals by being envious. In addition, individuals with lower CRT scores are also significantly more likely to exhibit a non-positive compassion parameter ($\beta \leq 0$), which in combination with envy ($\alpha > 0$) would be a sign of *antisocial*, spiteful motivations. As previously argued, combining both α and β is essential to obtain a complete picture of the motives driving social behavior. Our next analyses address this point.

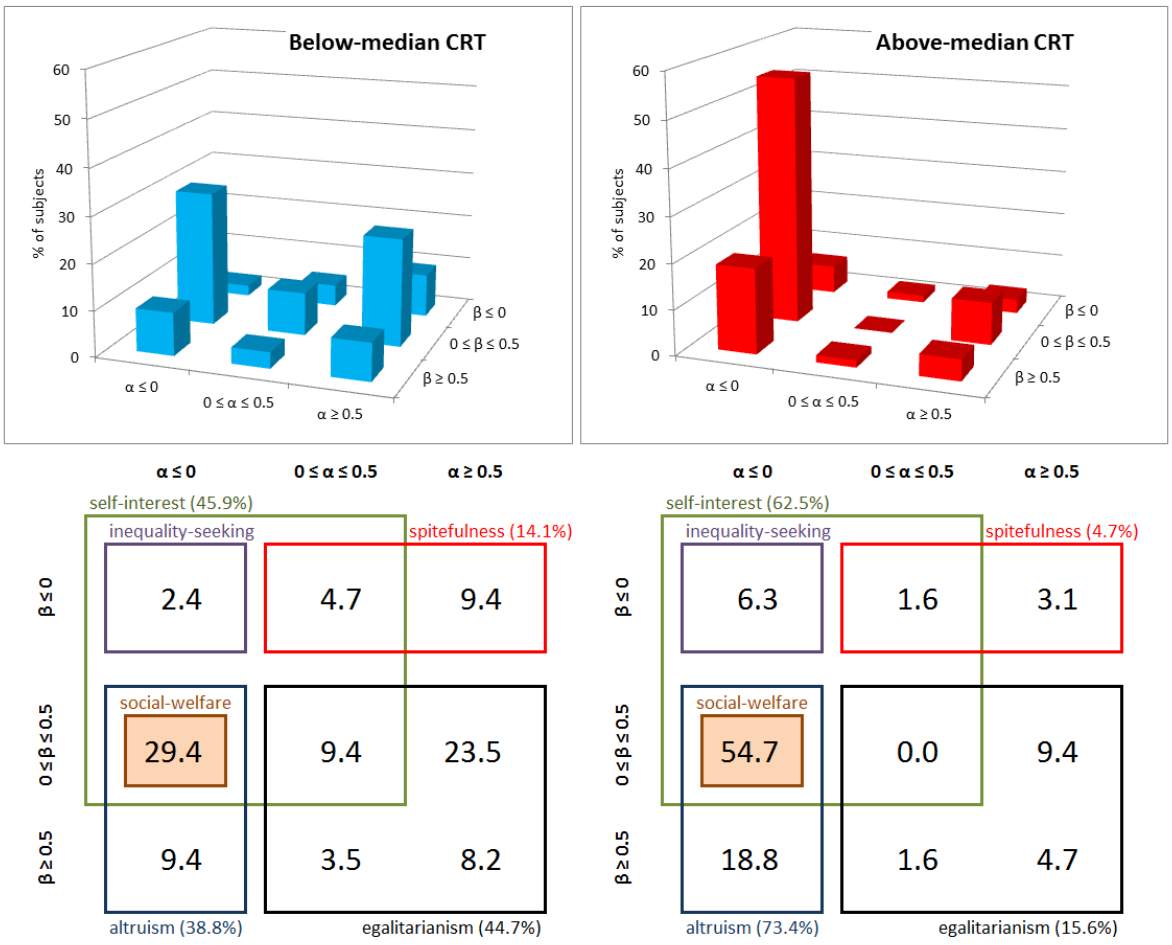

Figure 1. Classification of subjects according to the envy and compassion parameters, by CRT groups (Study 1). The figure displays the % of subjects that can be classified according to each combination of α and β and the social preferences which are consistent with each category, broken down into below-median (n = 85) and above-median (n = 64) CRT score groups.

Figure 1 displays the % of individuals who are classified according to all possible combinations of the α and β parameters. Note that we include only those subjects with consistent choices, that is, choices which lead to compatible estimates of both α and β. This procedure excludes only one subject (out of 150). The left and right panels refer to subjects with below- and above-median CRT scores. In the table below each 3D plot, we highlight which among the combinations of the α and β parameters are consistent with each of the six categories of social motives previously defined: altruism, social-welfare concerns, self-interest, egalitarianism, spitefulness and inequality-seeking. For instance, all the (α, β) categories that include the value of 0 for both parameters are consistent with self-interest. The four cells representing these categories are surrounded by a green line. Also, the two (α, β) categories that include negative values of α and

positive values of β are consistent with altruistic motives and are surrounded by a light blue line. As was suggested by the previous analyses, above-median CRTs are concentrated (55% of them) in the category "$\alpha \leq 0, \beta \in [0, 0.5]$", which is highlighted in Figure 1. The proportion of above-median CRT subjects belonging to this category is significantly larger than the proportion of above-median CRT subjects belonging to any other category (two-sided Normal Proportion tests, ps < 0.01). In the case of below-median CRT subjects a much lower proportion (29%) belong to the "$\alpha \leq 0, \beta \in [0, 0.5]$" category (two-sided Normal Proportion test, p < 0.01). This category is still the most populated category among below-median CRT individuals and the proportion of individuals belonging to this category is significantly larger than the proportion of below-median CRT individuals belonging to any other category (two-sided Normal Proportion tests, ps < 0.01) but the "$\alpha \geq 0.5, \beta \in [0, 0.5]$" category (p = 0.38).

The category "$\alpha \leq 0, \beta \in [0, 0.5]$" is consistent with both self-interest and altruism and can thus be seen as "weak altruism". Our choice of terminology is to refer as "weak" all the social preferences categories that are consistent with self-interest (i.e. $\alpha = 0$ and $\beta = 0$). We refer to as "strong" all the (α, β) social preferences categories which are not "weak". Note that the "weak altruism" category is also the only category that is consistent with social-welfare motives. In order to show that these subjects display a preference for social welfare, however, one must show that $-0.5 < \alpha < 0$ which cannot be demonstrated given the social preferences elicitation task used in this study.

In order to inquire further on the categorization of social preferences and highlight differences across CRT scores, we perform a multinomial Probit regression (see Figure 2). We estimate the likelihood that an individual is included in the category "$\alpha \leq 0, \beta \in [0, 0.5]$", as compared to each of the other eight categories. We include CRT scores and gender as regressors. In each cell representing an (α, β) category in Figure 2, we show the coefficient associated to CRT scores for the comparison of this specific (α, β) category with the default category ("$\alpha \leq 0, \beta \in [0, 0.5]$"), which is left empty. As expected, all the coefficients associated to CRT are negative, indicating that subjects with higher CRT scores are more likely to be included in the default category than in any of the other categories. These coefficients are highly significant (ps < 0.01) when comparing the default category with the following ones: "$\alpha \in [0, 0.5], \beta \leq 0$" (weakly spiteful),

"$\alpha \in [0, 0.5], \beta \in [0, 0.5]$" (weakly egalitarian) and "$\alpha \geq 0.5, \beta \leq 0$" (strongly spiteful). Marginal significance is also achieved when comparing the default option with the category "$\alpha \geq 0.5, \beta \geq 0.5$" (strongly egalitarian, p = 0.06). However, CRT scores are not statistically significant when comparing the default category with the remaining three categories (ps > 0.41): "$\alpha \leq 0, \beta \geq 0.5$" (strongly altruistic), "$\alpha \leq 0, \beta \leq 0$" (weakly inequality seeking) and "$\alpha \in [0, 0.5], \beta \geq 0.5$" (strongly egalitarian). Yet, the latter two categories contain only six and four observations, respectively. Finally, the coefficients of these three categories (-0.030, -0.011 and -0.117, respectively) are the only ones which significantly differ from that of the strongly spiteful category "$\alpha \geq 0.5, \beta \leq 0$" (p < 0.01, p < 0.01 and p = 0.06), which reports the highest coefficient in absolute value (-0.457).

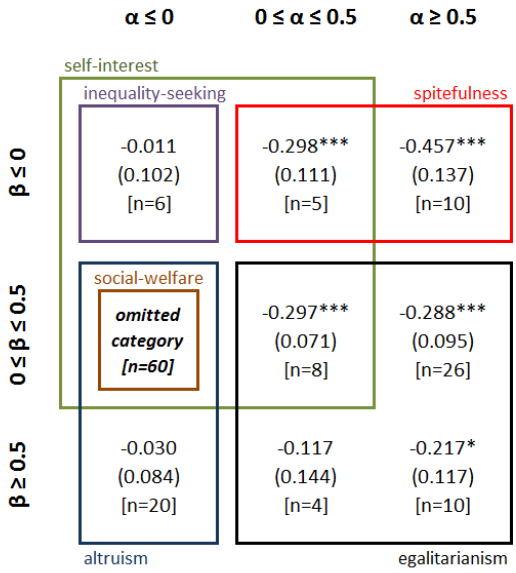

Figure 2. Output of multinomial Probit regression (Study 1). In each cell, the figure shows the coefficient of CRT obtained by comparing that specific social preference category with "$\alpha \leq 0, \beta \in [0, 0.5]$". *, **, *** denote significance at the 0.10, 0.05 and 0.01 level, respectively. Ll = -247.017, Wald χ^2 = 45.12 (p < 0.01), N = 149. Robust standard errors clustered on individuals are shown in parentheses, and the number of observations in square brackets.

Our classification thus suggests that high cognitive reflection is characteristic of individuals with $\alpha \leq 0$ and $\beta \in [0, 0.5]$, which corresponds to "weak" altruism, whereas less reflective individuals are more likely to be guided by either spiteful or egalitarian motives. Yet, our previous analysis of each of the four decisions in the social preferences elicitation task led to the more precise conclusion that high CRTs are characterized by $\alpha < 0$ and $\beta \in (0, 0.5)$. That is, high-CRT

individuals are unlikely to be purely selfish ($\alpha = 0$, $\beta = 0$); instead they can be considered as *mildly altruistic*.

Given the data of Study 1, high-CRT people are apparently more willing to give money to the other person than low-CRT people as long as it is not too costly for them to do so. Indeed, subjects with higher CRT scores are more willing to give money to the other person when it is costless (Decisions 1, 3 and 4) but not when it is very costly (Decision 2). Moreover, note that those subjects who give money to the other person in Decisions 1, 3 and 4 may respond to concerns for social welfare whereas such interpretation of giving is not valid for Decision 2.

However, substantial differences may still exist in the levels of envy (α) and compassion (β) among those subjects characterized as mildly altruistic. Some mildly altruistic individuals may be close to selfishness ($\alpha \approx 0$, $\beta \approx 0$) whereas others may not. Our data cannot separate these different types of subjects. To that end, we extend the social preferences elicitation task of Bartling et al. (2009). First, we include in our elicitation task a decision for which increasing the payoff of the other person above one's own is personally costly. This decision will allow us to isolate subjects who are practically selfish ($\alpha \approx 0$) in the negative domain of envy. Second, among mildly altruistic subjects there may be individuals with social-welfare concerns ($|\alpha_i|$, $|\beta_i| < 0.5$). To isolate people who care about social welfare, we need that increasing the *better-off* counterpart's payoff in the aforementioned decision also increases social welfare (i.e. the cost for the decision maker is lower than the increase in the other player's payoff). In addition, we need to include another decision for which increasing a *worse-off* counterpart's payoff at a personal cost also increases social welfare.

In order to dig into these issues and obtain a more refined assessment of the values of α and β, we thus modified the social-preferences task of Bartling et al. (2009) by adding two decisions which were designed along the lines of the previous discussion. This modified task was implemented in Study 2.

Study 2

Methods

Participants and general protocol. Participants were 158 students (51.90% female; mean age 21.52±2.63 (SD)) from the University Carlos III of Madrid in Spain. These participants were recruited from a database of more than 2,500 students. We conducted a total of eight sessions, three with 18 and 20 participants each and two with 22 participants. On average, sessions lasted for 60 minutes. As in Study 1, all subjects completed the same tasks in the same order and the social preferences elicitation task was performed before the CRT. In between the social preferences task and the CRT, participants completed a series of unrelated tasks for about 15 minutes and had a break of 10 minutes. All participants in the experiments reported in this Study agreed to the Participation Rules and Privacy Policy when they registered to participate in experiments. Anonymity was always preserved (in agreement with Spanish Law 15/1999 on Personal Data Protection) by randomly assigning a numerical code to identify the participants in the system. No association was ever made between their real names and the results. As is standard in socio-economic experiments, no ethic concerns are involved other than preserving the anonymity of participants. No deception was used. This procedure was checked and approved by the department of Economics of the University Carlos III of Madrid; the institution hosting the experiments. At that time no official IRB was established at the university.

Cognitive style assessment. As in Study 1, participants completed the extended version of the CRT developed by Toplak et al. (2014). In Table S2, we display the % of subjects answering each question correctly, split by gender. Again, males scored higher on the test than females. The full distribution of correct answers by males (mean = 3.22±1.73) and females (mean = 2.18±1.35) is provided in Figure S2. The test was not incentivized.

Social preferences elicitation. Participants made six choices between two possible allocations of money between themselves and another anonymous participant with whom they were randomly matched. Similarly to Study 1, in each experimental session, two participants and one of the six decisions were selected at random for payment. The choice of the first participant in the selected decision was used to allocate payoffs between the two participants. All decisions were anonymous. The first four decisions used the exact same payoffs as in Bartling et al. (2009).

Decisions 5 (advantageous domain) and 6 (disadvantageous domain) were designed for this particular experiment in such a way that the decision maker could increase the payoff of the other participant by €6 at a €2 cost. Thus, the cost for the decision-maker is low relative to the increase of the other's payoff (i.e. a 1:3 cost-to-benefit ratio) so that giving also increases social welfare. The new task allows us to disentangle four subcategories of the "weak altruism" category of Study 1 ($\alpha \leq 0$, $\beta \in [0, 0.5]$) which was the most populated category and also the only one which was consistent with welfare concerns. In Study 2 and in contrast to Study 1, we could identify subjects exhibiting combinations of α and β that are consistent with social-welfare concerns but *not* with self-interest. We could thus distinguish between "weak" and "strong" preferences for social welfare.

All the allocation decisions are described in Table 3. Option A always yielded an even distribution of money (€10 to both the self and the other participant) whereas option B yielded uneven payoffs. For each decision, we show in parentheses the envy/compassion parameter associated to choosing the egalitarian and non-egalitarian options (i.e. options A and B) and in square brackets the proportion of subjects who chose each option. Note that the model parameters associated to Decisions 1-4 are the same as in Study 1, except for the fact that in Decision 4 the threshold for the envy parameter is now 0.125 instead of 0.5.

Decision #	Option A		Option B	
	self, other		self, other	
1	€10,€10	($\beta \geq 0$) [86%]	€10,€6	($\beta \leq 0$) [14%]
2	€10,€10	($\beta \geq 0.5$) [27%]	€16,€4	($\beta \leq 0.5$) [73%]
3	€10,€10	($\alpha \geq 0$) [42%]	€10,€18	($\alpha \leq 0$) [58%]
4	€10,€10	($\alpha \geq 0.125$) [30%]	€11,€19	($\alpha \leq 0.125$) [70%]
5	€10,€10	($\beta \geq 0.25$) [42%]	€12,€4	($\beta \leq 0.25$) [58%]
6	€10,€10	($\alpha \geq -0.25$) [84%]	€8,€16	($\alpha \leq -0.25$) [16%]

Table 3. Decisions in the social preferences task (Study 2). For each option, we display the payoff for the decision-maker and the recipient, the associated model parameters (in parentheses) and the % of subjects choosing it (in square brackets).

Results and discussion

Decision analysis

Table 4 reports the results of a series of Probit regressions where the choice of option B in each decision is regressed as a function of CRT scores, controlling for gender. Robust standard errors clustered on individuals are presented in parentheses. In Figure S4, we display the proportion of subjects choosing option B in each decision, for individuals with below-median (i.e. two or less correct answers, n = 85) and above-median (n = 73) CRT scores.

We find that CRT is negatively and (marginally) significantly related to choosing option B in Decision 1 (p = 0.09), indicating that individuals with higher CRT scores are less likely to display $\beta \leq 0$, in line with the findings in Study 1. Also, the qualitative nature of our statistical results is robust to using a binary categorization of CRT instead of CRT scores. The effect of cognitive reflection on Decision 1 is actually more significant (p = 0.04, Table S4, column 1) when the binary categorization is used as explanatory variable. As in Study 1, a two-sided binomial test rejects the hypothesis that individuals are indifferent between the two options (for both below- and above-median CRT scores, the proportion of subjects choosing option B is well below 50%; ps < 0.01). That is, the majority of subjects, especially those with higher CRT scores, seem to display $\beta > 0$.

In Decision 2, we observe some discrepancy with respect to Study 1 where CRT was not a significant predictor. In Study 2, CRT scores are positively and significantly related to choosing option B, indicating that higher CRT individuals are more likely to exhibit $\beta \leq 0.5$. Yet, this relationship is only marginally significant (p = 0.09) and even turns insignificant when the binary categorization of CRT is used (p = 0.32, Table S4, column 2). As in Study 1, the proportion of subjects choosing option B in Decision 2 is higher than 50% in both CRT groups (two-sided binomial tests, ps < 0.01; see Figure S4). That is, the majority of subjects, especially those with higher CRT scores, seem to be characterized by $\beta < 0.5$.

With regards to Decisions 3 and 4, the results are similar to those of Study 1. Specifically, CRT is positively and (marginally) significantly associated with the choice of option B in Decision 3 (p = 0.07), implying $\alpha \leq 0$. This relationship turns significant at the 5% level when the binary

CRT variable is used (p = 0.03, Table S4, column 3). From Figure S4, we see that roughly 48% of below-median CRT subjects choose option B in Decision 3 (which is not significantly different from 50%, two-sided binomial test, p = 0.66, so we cannot reject that they are, on average, indifferent between both options: $\alpha = 0$). In contrast, 68% of above-median CRT subjects choose option B (which is significantly different from 50%, p < 0.01). So, high-CRT subjects seem to display $\alpha < 0$. In Decision 4, the choice of option B is positively and significantly predicted by CRT (p = 0.04; the binary CRT categorization yields p = 0.05, Table S4, column 4), implying that higher CRT subjects are more likely to display $\alpha \leq 0.125$. Indeed, about 79% of above-median CRT subjects choose option B in Decision 4 (Figure S4), which is significantly different from 50% (two-sided binomial test, p < 0.01), whereas 61% of below-median CRT subjects did so (which is also significantly different from 50%, p = 0.05). Thus, $\alpha <$ 0.125 seems to best characterize the majority of subjects, especially those with high CRT scores.

Dep var:	Decision 1 $\beta \leq 0$ (vs ≥ 0) (1)	Decision 2 $\beta \leq 0.5$ (vs ≥ 0.5) (2)	Decision 3 $\alpha \leq 0$ (vs ≥ 0) (3)	Decision 4 $\alpha \leq 0.125$ (vs ≥ 0.125) (4)	Decision 5 $\beta \leq 0.25$ (vs ≥ 0.25) (5)	Decision 6 $\alpha \leq -0.25$ (vs ≥ -0.25) (6)
CRT	-0.149*	0.122*	0.124*	0.150**	0.109	-0.044
	(0.088)	(0.073)	(0.069)	(0.072)	(0.067)	(0.084)
female	0.306	0.070	-0.285	-0.462**	0.294	-0.146
	(0.276)	(0.227)	(0.213)	(0.223)	(0.216)	(0.257)
cons	-0.898***	0.252	0.017	0.390	-0.250	-0.812***
	(0.334)	(0.271)	(0.262)	(0.264)	(0.260)	(0.315)
ll	-60.929	-90.969	-103.938	-90.902	-105.927	-68.750
Wald χ^2	6.18**	2.89	6.86**	10.74***	3.35	0.44
pseudo R^2	0.045	0.017	0.035	0.063	0.016	0.004

Table 4. Non-egalitarian choice (option B) as a function of CRT (Study 2). Probit estimates. The α and β parameters associated with the dependent variable are displayed on top of each column. Robust standard errors clustered on individuals are shown in parentheses. *, **, *** denote significance at the 0.10, 0.05 and 0.01 level, respectively. $N = 158$ in all regressions.

In Decision 5, CRT does not yield a significant effect (p = 0.11; using the binary CRT variable, p = 0.39, Table S4, column 5). While 60% of above-median CRT subjects choose option B in Decision 5 (this proportion is marginally significantly different from 50%, two-sided binomial test, p = 0.10), this percentage shrinks to 55% for below-median CRT subjects (not significantly different from 50%, p = 0.38) (Figure S4). This indicates that most high-CRT subjects are characterized by $\beta < 0.25$, whereas the median β seems to be close to 0.25 for low-CRT subjects.

Finally, in Decision 6, where option B implies $\alpha \leq -0.25$, the coefficient associated to CRT is far from significant (p = 0.60; also using the binary CRT variable, p = 0.74, Table S4, column 6). About 16% of below-median CRT subjects and 15% of above-median CRT subjects choose option B in Decision 6 (both proportions are significantly different from 50%, two-sided binomial tests, ps < 0.01; see Figure S4), which implies that the majority of subjects is best characterized by $\alpha > -0.25$, regardless of CRT scores.

In sum, the previous analysis suggests that high-CRT individuals are best described by $\alpha \in (-0.25, 0)$ and $\beta \in (0, 0.25)$, whereas the distribution of the envy and compassion parameters of low CRT subjects is much more disperse.

Social preferences categorization

Now, we proceed by categorizing each individual according to their social preferences.

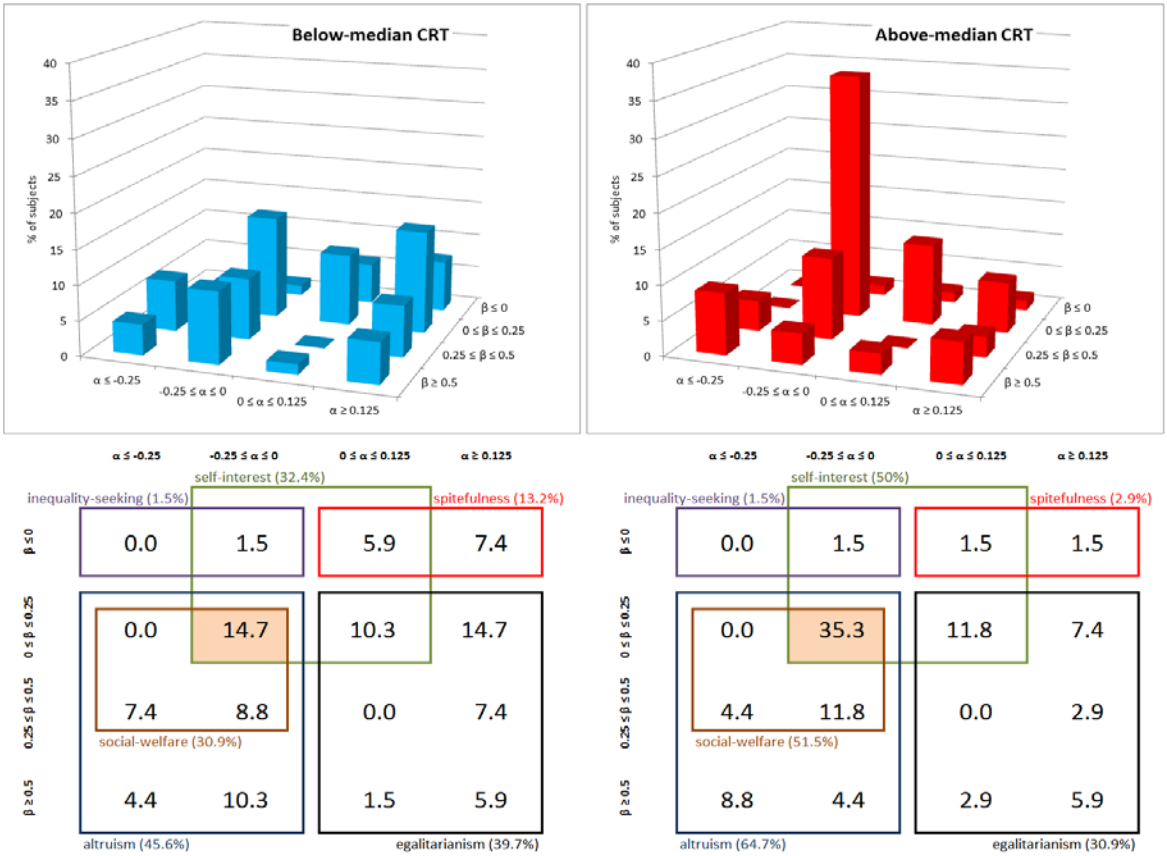

Figure 3. Classification of subjects according to the envy and compassion parameters, by CRT groups (Study 2). The figure displays the % of subjects that can be classified according to each combination of α and β and the social preferences which are consistent with each category, broken down into below-median (n = 68) and above-median (n = 68) CRT score groups.

In Figure 3, we display the proportion of subjects that are characterized by each of the sixteen combinations of the envy and compassion parameters. We represent below-median CRT subjects on the left panel and above-median CRT subjects on the right panel. In our social preferences categorization, we excluded 22 subjects whose choices were inconsistent, so we ended up with 136 observations (68 below-median and 68 above-median CRT subjects). No individuals were assigned to the following categories: "$\alpha \leq$ -0.25, $\beta \leq 0$" (strongly inequality seeking), "$\alpha \leq$ -0.25, $\beta \in$ [0, 0.25]" (strongly altruistic with social-welfare concerns) and "$\alpha \in$ [0, 0.125], $\beta \in$ [0.25, 0.5]" (strongly egalitarian).

As expected, above-median CRT subjects are concentrated (35% of them) in the category "$\alpha \in$ [-0.25, 0], $\beta \in$ [0, 0.25]", which again represents "weak altruism", whereas below-median CRT subjects are more dispersed across categories, similarly to Study 1. The proportion of above-median CRT subjects belonging to this category is significantly larger than the proportion of above-median CRT subjects belonging to any other category (two-sided Normal Proportion test, ps < 0.01). In the case of below-median CRT subjects a much lower proportion of people (15%) belong to the "$\alpha \leq 0$, $\beta \in$ [0, 0.5]" category ($p < 0.01$). This category is still the most populated category among below-median CRT subjects but the proportion of below-median CRT subjects belonging to this category is only significantly larger than six out of the fifteen other categories. Note that, in contrast to Study 1 where there was only one category consistent with social-welfare concerns, Study 2 allows us to identify different degrees of such concerns. The category defining the majority of above-median CRT subjects ("$\alpha \in$ [-0.25, 0], $\beta \in$ [0, 0.25]") corresponds to "weak" social-welfare concerns.

In order to further explore these observations, we conducted a multinomial Probit regression, the results of which are presented in Figure 4. As for Study 1, CRT and gender are used as regressors. The most populated category, "$\alpha \in$ [-0.25, 0], $\beta \in$ [0, 0.25]", is used as the default category for the regression analysis. The numbers inside the remaining cells indicate the effect of CRT score on the likelihood that an individual is included in this specific category as compared to the default category. As expected, all estimates are negative indicating that subjects with higher CRT scores are more likely to belong to the default category "$\alpha \in$ [-0.25, 0], $\beta \in$ [0, 0.25]" than to the remaining categories. The effect of CRT is statistically significant when comparing

the default category to the following ones: "$\alpha \in [-0.25, 0]$, $\beta \geq 0.5$" (strongly altruistic, $p = 0.02$), "$\alpha \in [0, 0.125]$, $\beta \leq 0$" (weakly spiteful, $p < 0.01$), "$\alpha \geq 0.125$, $\beta \in [0.25, 0.5]$" (strongly egalitarian, $p < 0.01$) and "$\alpha \geq 0.125$, $\beta \geq 0.5$" (strongly egalitarian, $p < 0.01$). The effect of CRT is marginally significant with respect to "$\alpha \geq 0.125$, $\beta \leq 0$" (strongly spiteful, $p = 0.08$) and "$\alpha \geq 0.125$, $\beta \in [0, 0.25]$" (strongly egalitarian, $p = 0.08$) and close to significance with respect to "$\alpha \leq -0.25$, $\beta \in [0.25, 0.5]$" (strongly altruistic with social-welfare concerns, $p = 0.13$). The five remaining categories did not yield significant CRT effects (ps > 0.23). An interesting difference of Study 2 with respect to Study 1 is that two of the "strong altruism" categories show significant (or nearly significant) differences with the default group. This did not happen in Study 1, where there was only one such category (namely "$\alpha \leq 0$, $\beta \geq 0.5$") for which the associated coefficient was largely insignificant. Note that here the strongest difference is given by the comparison with the following category "$\alpha \geq 0.125$, $\beta \in [0.25, 0.5]$" (strongly egalitarian), which is the category higher CRT subjects are less likely to belong to. However, the coefficient associated to this category only differs significantly from the coefficient of the following categories: "$\alpha \in [-0.25, 0]$, $\beta \leq 0$" (weakly inequality seeking; note that only two subjects belong to this category), "$\alpha \in [0, 0.125]$, $\beta \in [0, 0.25]$" (weakly egalitarian) and "$\alpha \in [0, 0.125]$, $\beta \geq 0.5$" (strongly egalitarian; only three subjects belong to this category) (ps < 0.05). The coefficient associated to the "$\alpha \geq 0.125$, $\beta \in [0.25, 0.5]$" category marginally differs from the coefficients of the following categories: "$\alpha \leq -0.25$, $\beta \geq 0.5$" (strongly altruistic) and "$\alpha \in [-0.25, 0]$, $\beta \in [0.25, 0.5]$" (strongly altruistic with social-welfare concerns) (ps = 0.08).

Taken together, the results of Study 2 indicate that high cognitive reflection is characteristic of individuals who make choices consistent with mildly altruistic motives that increase social welfare at a *very low* cost. Low cognitive reflection is characteristic of individuals who make decisions consistent with either egalitarian or spiteful motives. These findings are consistent with Study 1. In slight contrast to Study 1, however, low-CRT people are also associated with strong altruistic motivations. It is important to note that, in Study 2, we were able to split the weak altruism/social-welfare preferences category into four subcategories. In contrast to Study 1, we could therefore isolate strong social-welfare concerns from weak social-welfare concerns and conclude that it is the latter which best characterizes individuals with high CRT scores. This methodological feature of Study 2 may thus have facilitated the observation of a difference in

terms of CRT scores between those subjects included in the default category and those classified as strongly altruistic.

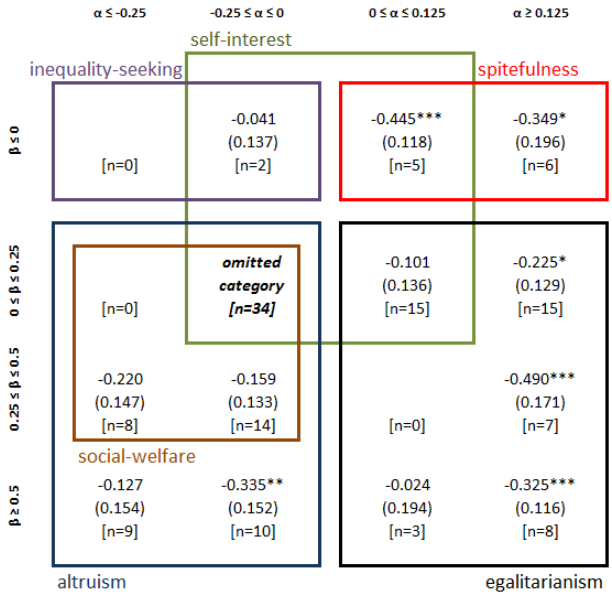

Figure 4. Output of multinomial Probit regression (Study 2). In each cell, the figure shows the coefficient of CRT obtained by comparing that specific social preference category with "$\alpha \in [-0.25, 0]$, $\beta \in [0, 0.25]$". *, **, *** denote significance at the 0.10, 0.05 and 0.01 level, respectively. Ll = -300.325, Wald $\chi^2 = 50.17$ ($p < 0.01$), N = 136. Robust standard errors clustered on individuals are shown in parentheses, and the number of observations in square brackets.

General discussion

Cognitive reflection and social preferences: our insights

In two studies, we showed that those individuals with a more reflective cognitive style (i.e. those who are less likely to rely on intuitive, System 1 processes) are more likely to make choices consistent with mildly altruistic motives in simple monetary decisions free of strategic and reciprocal concerns. These results suggest that behaviors that increase social welfare by increasing others' payoffs at a very low or no cost for the individual may be the result of conscious deliberation rather than automatic heuristics. Behaviors driven by egalitarian or spiteful concerns, however, appear to be more intimately associated with intuition.[6]

[6] In a related strand of research on the psychological underpinnings of social behavior, Espín et al. (2012, 2013) have shown that those individuals who discount the future more heavily display more spite-based but *not* egalitarian behavior in economic games (namely in a dual-role ultimatum game and a social dilemma game with punishment). Since we find that lower CRT scores are related to both spitefulness and egalitarianism, our results thus support Espín et al.'s arguments that the social behaviors associated with short-run (vs. long-run) goals differ from those

While the above findings are robust across the two studies, we also find a slight but remarkable difference with respect to strongly altruistic choices that increase the other's payoff at a relatively high cost to the individual. In Study 1 reflective subjects were quite likely to make such choices whereas in Study 2 they were not. Although methodological differences across studies (in Study 2 weak altruism was divided into four subcategories and strong altruism into two subcategories) may have facilitated the observation of this divergence, it might also have been influenced by cultural differences (Study 1 was conducted in the US while Study 2 was conducted in Spain). This difference may also be explained by the existence of ceiling effects as the average level of cognitive reflection, as measured by the number of correct answers to the CRT, was higher (25% higher, two-sided t-test: $p < 0.01$) in Study 1. Exploring these possibilities is an interesting avenue for future research.

Toward reconciliation: a unified view of the cognitive basis of social behavior

The Social Heuristics Hypothesis

At first sight, it might seem that more reflective individuals are guided by "weaker" social motivations as they are typically less likely to be classified in the categories representing strong social preferences. Accordingly, it may be tempting to interpret our findings as evidence that cognitive reflection goes along with self-interest in (non-strategic) one-shot social interactions. This would be, however, an incorrect interpretation of our findings because self-interest cannot explain why the most reflective individuals are overwhelmingly characterized as "mildly" altruistic while not being affected by other social preferences like spitefulness or egalitarianism. Therefore, it is not self-interest per se but a very particular mixture of self-interest and altruistic/social-welfare concerns that characterizes reflective individuals. In terms of the parameters of the generalized version of the Fehr-Schmidt's (1999) model used here, high cognitive reflection is associated with a combination of slightly negative values of envy (α) and slightly positive values of compassion (β). Similar results have been obtained through structural estimation of the individuals' envy and compassion parameters in Ponti & Rodríguez-Lara (2015). Moreover, there are much less individual differences in these parameters among

associated with intuitive (vs. controlled) processes (even though temporal discounting is negatively correlated with CRT scores; see Frederick et al. 2005, Bosch-Domènech et al. 2014).

individuals with high CRT scores than among individuals with low CRT scores. While mean values of envy appear to be higher for individuals with a less reflective cognitive style, the relationship between CRT scores and compassion is more complex. Indeed, either high or very low (even negative) values of β can be associated with low cognitive reflection. Thus, we would not have been able to uncover some of the key differences between groups if we had focused on estimating mean values of the model parameters.

From the viewpoint of the Social Heuristics Hypothesis (Rand et al. 2014b), our results suggest that behaviors driven by either egalitarianism or spitefulness (and possibly strong altruism) may be internalized as heuristics, which ultimately implies that they may be, on average, advantageous in daily-life interactions. Indeed, neurobiological research indicates that humans experience psychological satisfaction from observing equitable outcomes (Tricomi et al. 2010, Zaki & Mitchell 2011) but also from out-earning others (Fliessbach et al. 2007, Bault et al. 2011), even if their own absolute payoff is unaffected. On the other hand, reflection should lead people to adapt their decision rules to the environment at hand (e.g. Kahneman 2011). Under this logic, the present results indicate that the most adaptive decisions in one-shot, non-strategic social interactions are those guided by mildly altruistic motives.

These findings can shed light on the current debate regarding whether (pro)social behavior is automatic or deliberate (Rand & Nowak 2013, Zaki & Mitchel 2013). Previous research has led to ostensibly contradictory results which have partly been accounted for by the existence of moderator variables (e.g. subjects' prior experience in economic experiments; Cone & Rand 2014, Rand et al. 2014a, 2014b) and confounding factors (linked, for example, to the use of reaction times to infer the effect of reflection on behavior; Evans et al. 2014, Recalde et al. 2014). Yet, our findings reveal that another non-negligible portion of these apparently conflicting findings can be reconciled by accounting for two often-ignored factors. First, different motives can lead to identical choices in the experimental set-ups normally used to infer the nature of social behavior (Charness & Rabin 2002). Second, by putting the focus almost exclusively on the conflict between prosociality and self-interest, previous research has tended to overlook antisocial motivations that can trigger behaviors which may appear as selfish or even prosocial (e.g. Espín et al. 2012, Brañas-Garza et al. 2014).

Revisiting previous findings in standard economic games

We start by considering the standard dictator game, where the decision maker has to split a sum of money between herself and a passive recipient (Forsythe et al. 1994). Since giving in the dictator game is typically the result of either strongly egalitarian or strongly altruistic motives, our findings suggest that it should be an intuitive decision (Cornelissen et al. 2011, Ruff et al. 2013, Cappelen et al. 2014). However, since giving nothing can be the result not only of self-interest but also of spitefulness, it may also respond to intuition (Martinsson et al. 2012, Xu et al. 2012, Achtziger et al. 2015). The exact proportion of these different motives in a specific population would thus dramatically influence whether intuitive decision making is linked to more or less generosity in the dictator game.

The ultimatum game (Güth et al. 1982) is identical to the dictator game except for the fact that the second player (the responder) can either accept or reject the proposed distribution of money. If she accepts, the proposal is implemented accordingly; in case of rejection, both players earn nothing. The behavior of the first player (the proposer) is strongly influenced by strategic concerns (avoidance of rejection) so the intrinsic motivations behind proposers' behavior are difficult to uncover (e.g. Prasnikar & Roth 1992). By contrast, responders' behavior is not shaped by strategic concerns and the decision to reject a low offer has been primarily rationalized as the result of the individual's aversion to unequal outcomes (Fehr & Schmidt 1999). In this vein, the responder's decisions are often interpreted as a conflict between the money-maximizing strategy of accepting an unfair but positive offer and the egalitarian, prosocial preference to reject it. Recent research has shown, however, that altruistic, prosocial motives can also lead to acceptance (Staffiero et al. 2013) and that spiteful, antisocial motives can also lead to rejection (Brañas-Garza et al. 2014). According to these observations, our results suggest that rejections should be driven by intuition (whether they are guided by egalitarian or spiteful motivations) whereas acceptance should be driven by reflection in most (mild altruism), but not all (strong altruism) cases. The fact that rejection [acceptance] is guided by intuitive [deliberative] processes is consistent with the majority of previous findings (e.g. Sutter et al. 2003, Tabibnia et al. 2008, Grimm & Mengel 2011), although there exist some challenging neurobiological evidence (Knoch et al. 2006, 2010).

With regards to social dilemma games, our results suggest that cooperation can be the result of conscious deliberation (e.g. Lohse et al. 2014, Martinsson et al. 2014) if it follows from mildly altruistic motives but can also be intuitive (e.g. Rand et al. 2012, Nielsen et al. 2014) when it is guided either by a desire to equalize payoffs with expectedly cooperative partners or by a strongly altruistic desire to unconditionally increase others' payoffs. Likewise, defection can be driven by reflection if the cost of increasing others' payoffs is sufficiently high so that only highly altruistic but not mildly altruistic individuals are willing to cooperate but can also be driven by intuition if it is the result of spiteful or egalitarian (i.e. to equalize payoffs with expectedly uncooperative partners) motives. Whether cooperation is (statistically) an intuitive or reflective decision in social dilemma experiments could thus depend on the distribution of social motives in the sample under scrutiny.

In sum, our findings highlight that the analysis of the cognitive basis of social behavior is likely to be more complex than previously thought. It must also be said, however, that strategic issues and reciprocity (which were voluntarily absent of our study) may play an essential role in social dilemma and ultimatum games (e.g. Charness & Rabin 2002, Falk & Fischbacher 2006, Fehr & Schmidt 2006), blurring further the analysis of the cognitive basis of social behavior.

Finally, our results are based on a trait approach to cognitive reflection and it would thus be important for future research to assess the robustness of these findings to experimental manipulations of intuitive processing. Identifying the neurobiological underpinnings of these individual differences in trait reflectiveness and their relationship to social preferences appears as a necessary next step towards achieving a more complete understanding of the cognitive basis of human social behavior (Nash et al. 2015).

Acknowledgments

The authors acknowledge financial support from the International Foundation for Research in Experimental Economics, the Argyros School of Business and Economics at Chapman University, the Spanish Ministry of Education [Grant 2012/00103/001], Spanish Plan Nacional I+D MCI [ECO2013-44879-R], 2014-17, and Proyectos de Excelencia de la Junta Andalucía [P12.SEJ.1436], 2014-18. This research started when the third author was working at the University of Granada.

References

Abbink, K., Brandts, J., Herrmann, B., & Orzen, H. (2010). Intergroup conflict and intra-group punishment in an experimental contest game. *The American Economic Review, 100*(1), 420-447.

Bartling, B., Fehr, E., Maréchal, M. A., & Schunk, D. (2009). Egalitarianism and competitiveness. *The American Economic Review, 99*(2), 93-98.

Bault, N., Joffily, M., Rustichini, A., & Coricelli, G. (2011). Medial prefrontal cortex and striatum mediate the influence of social comparison on the decision process. *Proceedings of the National Academy of Sciences of the USA, 108*(38), 16044-16049.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *The American Economic Review, 90*(1), 166-193.

Bosch-Domènech, A., Brañas-Garza, P., & Espín, A. M. (2014). Can exposure to prenatal sex hormones (2D: 4D) predict cognitive reflection?. *Psychoneuroendocrinology, 43*, 1-10.

Bowles, S., & Gintis, H. (2003). Origins of human cooperation. In Hammerstein, P. (Ed.), *Genetic and Cultural Evolution of Cooperation*, pp. 429-443.

Brañas-Garza, P., Espín, A. M., Exadaktylos, F., & Herrmann, B. (2014). Fair and unfair punishers coexist in the Ultimatum Game. *Scientific Reports, 4*, 6025.

Brañas-Garza, P., Garcia-Muñoz, T., & González, R. H. (2012a). Cognitive effort in the beauty contest game. *Journal of Economic Behavior & Organization, 83*(2), 254-260.

Brañas-Garza, P., Meloso, D., & Miller, L. (2012b). Interactive and moral reasoning: A comparative study of response times. *University of Bocconi Working Paper*, 440.

Camerer, C. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Capraro, V. (2013). A model of human cooperation in social dilemmas. *PloS One, 8*(8), e72427.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics, 117*(3), 817-869.

Cone, J., & Rand, D. G. (2014), Time pressure increases cooperation in competitively framed social dilemmas. *Plos One 9(*12), e115756.

Corgnet, B., Hernán-González, R., Kujal, P., & Porter, D. (2014a). The effect of earned versus house money on price bubble formation in experimental asset markets. *Review of Finance*, rfu031.

Corgnet, B., Hernán-González, R., & Mateo, R. (2014b). Cognitive reflection and the diligent worker: An experimental study of millennials, mimeo Chapman University.

Cornelissen, G., Dewitte, S., & Warlop, L. (2011). Are social value orientations expressed automatically? Decision making in the dictator game. *Personality and Social Psychology Bulletin*, *37*(8), 1080-1090.

Cox, C. A. (2013). Inequity aversion and advantage seeking with asymmetric competition. *Journal of Economic Behavior & Organization*, *86*, 121-136.

Cox, J. C., Friedman, D., & Gjerstad, S. (2007). A tractable model of reciprocity and fairness. *Games and Economic Behavior*, *59*(1), 17-45.

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences of the USA*, *111*(48), 17320-17325.

Cueva, C., Iturbe-Ormaetxe, I., Mata-Perez, E., Ponti, G., Sartarelli, M., Yu, H., & Zhukova, V. (2015). Cognitive (ir)reflection: New experimental evidence, mimeo Universidad de Alicante.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology*, *31*(1), 169-193.

DeWall, C. N., Baumeister, R. F., Gailliot, M. T., & Maner, J. K. (2008). Depletion makes the heart grow less helpful: Helping as a function of self-regulatory energy and genetic relatedness. *Personality and Social Psychology Bulletin*, *34*(12), 1653-1662.

Duffy, S., & Smith, J. (2014). Cognitive load in the multi-player prisoner's dilemma game: Are there brains in games? *Journal of Behavioral and Experimental Economics*, *51*, 47-56.

Engelmann, D. (2012). How not to extend models of inequality aversion. *Journal of Economic Behavior & Organization*, *81*(2), 599-605.

Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *The American Economic Review*, *94*(4), 857-869.

Espín, A. M., Brañas-Garza, P., Herrmann, B., & Gamella, J. F. (2012). Patient and impatient punishers of free-riders. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1749), 4923-4928.

Espín, A. M., Exadaktylos, F., Herrmann, B., & Brañas-Garza, P. (2013). Short-and long-run goals in ultimatum bargaining. *ESI Chapman Working Papers*, 13-17.

Evans, A. M., Dillon, K. D., & Rand, D. G. (2014). Reaction times and reflection in social dilemmas: extreme responses are fast, but not intuitive. Available at SSRN: http://ssrn.com/abstract=2436750

Exadaktylos, F., Espín, A. M., & Brañas-Garza, P. (2013). Experimental subjects are not different. *Scientific Reports*, *3*, 1213.

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, *73*(6), 2017-2030.

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, *54*(2), 293-315.

Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: The neural circuitry of social preferences. *Trends in Cognitive Sciences*, *11*(10), 419-427.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137-140.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*(3), 817-868.

Fehr, E., & Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism–experimental evidence and new theories. In Kolm, S-C, Ythier, J. M. (Eds.), *Handbook of the Economics of Giving, Altruism and Reciprocity, Vol. 1*, pp. 615-691.

Fliessbach, K., Weber, B., Trautner, P., Dohmen, T., Sunde, U., Elger, C. E., & Falk, A. (2007). Social comparison affects reward-related brain activity in the human ventral striatum. *Science*, *318*(5854), 1305-1308.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, *6*(3), 347-369.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25-42.

Fromell, H., Nosenzo, D., & Owens, T. Tradeoffs between self-interest and other-regarding preferences cause willpower depletion. *CeDEx Discussion Papers*, 2014-14.

Fudenberg, D., & Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica: Journal of the Econometric Society*, *54*(3), 533-554.

Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1518), 791-806.

Gächter, S., & Herrmann, B. (2011). The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia. *European Economic Review*, *55*(2), 193-210.

Gardner, A., & West, S. A. (2004). Spite and the scale of competition. *Journal of evolutionary biology*, *17*(6), 1195-1203.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, *206*(2), 169-179.

Grimm, V., & Mengel, F. (2011). Let me sleep on it: Delay reduces rejection rates in ultimatum games. *Economics Letters*, *111*(2), 113-115.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367-388.

Gutiérrez-Roig, M., Gracia-Lázaro, C., Perelló, J., Moreno, Y., & Sánchez, A. (2014). Transition from reciprocal cooperation to persistent behaviour in social dilemmas at the end of adolescence. *Nature Communications*, *5*, 4362.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology*, *7*(1), 1-52.

Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science*, *316*(5831), 1622-1625.

Hauge, K. E., Brekke, K. A., Johansson, L. O., Johansson-Stenman, O., & Svedsäter, H. (2014). Keeping others in our mind or in our heart? Distribution games under cognitive load. *University of Gothenburg Working Papers in Economics*, 600.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *The American Economic Review*, *91*(2), 73-78.

Herrmann, B., & Orzen, H. (2008). The appearance of homo rivalis: Social preferences and the nature of rent seeking. *CeDEx Discussion Paper Series*, 2008-10.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362-1367.

Hogarth, R. M. (2001). *Educating Intuition*. University of Chicago Press.

Jaber-López, T., García-Gallego, A., Perakakis, P., & Georgantzis, N. (2014). Physiological and behavioral patterns of corruption. *Frontiers in Behavioral Neuroscience*, *8*, 434.

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y. F., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning – implications for training and transfer. *Intelligence*, *38*(6), 625-635.

Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1553), 2635-2650.

Jensen, K. (2012). Who cares? Other-regarding concerns—decisions with feeling. In Hammerstein, P., Stevens, J. R. (Eds.), *Evolution and the Mechanisms of Decision Making*, pp. 299-317.

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

Kimbrough, E. O., & Reiss, J. P. (2012). Measuring the distribution of spitefulness. *Plos One*, *7*(8), e41812.

Kirchsteiger, G. (1994). The role of envy in ultimatum games. *Journal of Economic Behavior & Organization*, *25*(3), 373-389.

Knoch, D., Gianotti, L. R., Baumgartner, T., & Fehr, E. (2010). A neural marker of costly punishment behavior. *Psychological Science, 21*(3), 337-342.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, *314*(5800), 829-832.

Ledyard, J. (1995). Public goods: A survey of experimental research. In Kagel, J., Roth, A. (Eds.), *The Handbook of Experimental Economics*, Princeton University Press, pp. 111-194.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, *1*(3), 593-622.

Loewenstein, G. F., & O'Donoghue, T. (2004). Animal spirits: Affective and deliberative processes in economic behavior. *Available at SSRN 539843*.

Loewenstein, G. F., Thompson, L., & Bazerman, M. H. (1989). Social utility and decision making in interpersonal contexts. *Journal of Personality and Social psychology, 57*(3), 426.

Lohse, J., Goeschl, T., & Diederich, J. (2014). Giving is a question of time: Response times and contributions to a real world public good. *University of Heidelberg Department of Economics Discussion Paper Series*, 566.

López-Pérez, R. (2008). Aversion to norm-breaking: A model. *Games and Economic Behavior*, *64*(1), 237-267.

Lotito, G., Migheli, M., & Ortona, G. (2013). Is cooperation instinctive? Evidence from the response times in a public goods game. *Journal of Bioeconomics*, *15*(2), 123-133.

Martinsson, P., Myrseth, K. O. R., & Wollbrant, C. (2012). Reconciling pro-social vs. selfish behavior: On the role of self-control. *Judgment and Decision Making*, *7*(3), 304-315.

Martinsson, P., Myrseth, K. O. R., & Wollbrant, C. (2014). Social dilemmas: When self-control benefits cooperation. *Journal of Economic Psychology, 45*, 213-236.

McCall, C., Steinbeis, N., Ricard, M., & Singer, T. (2014). Compassion meditators show less anger, less punishment, and more compensation of victims in response to fairness violations. *Frontiers in Behavioral Neuroscience, 8*, 424.

McCullough, M. E., Pedersen, E. J., Schroder, J. M., Tabak, B. A., & Carver, C. S. (2012). Harsh childhood environmental characteristics predict exploitation and retaliation in humans. *Proceedings of the Royal Society B: Biological Sciences*, rspb20122104.

Moore, D. A., & Loewenstein, G. (2004). Self-interest, automaticity, and the psychology of conflict of interest. *Social Justice Research, 17*(2), 189-202.

Nash, K., Gianotti, L. R., & Knoch, D. (2015). A neural trait approach to exploring individual differences in social preferences. *Frontiers in Behavioral Neuroscience, 8*, 458.

Nielsen, U. H., Tyran, J. R., & Wengström, E. (2014). Second thoughts on free riding. *Economics Letters, 122*(2), 136-139.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science, 314*(5805), 1560-1563.

Oechssler, J., Roider, A., & Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization*, *72*(1), 147-152.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*(1), 163-177.

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition, 123*(3), 335-346.

Perc, M., & Szolnoki, A. (2008). Social diversity and promotion of cooperation in the spatial prisoner's dilemma game. *Physical Review E, 77*(1), 011904.

Peysakhovich, A., Nowak, M. A., & Rand, D. G. (2014). Humans display a 'cooperative phenotype' that is domain general and temporally stable. *Nature Communications, 5*, 4939.

Peysakhovich, A., & Rand, D. G. (2015). Habits of virtue: creating norms of cooperation and defection in the laboratory. *Management Science, In press*.

Piovesan, M., & Wengström, E. (2009). Fast or fair? A study of response times. *Economics Letters, 105*(2), 193-196.

Ponti G., & Rodríguez-Lara, I. (2015). Social Preferences and Cognitive Reflection: Evidence from Dictator Game Experiment. mimeo.

Prasnikar, V., & Roth, A. E. (1992). Considerations of fairness and strategy: Experimental data from sequential games. *The Quarterly Journal of Economics*, *107*(3), 865-888.

Prediger, S., Vollan, B., & Herrmann, B. (2014). Resource scarcity and antisocial behavior. *Journal of Public Economics*, *119*, 1-9.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, *83*(5), 1281-1302.

Raihani, N. J. (2014). Hidden altruism in a real-world setting. *Biology Letters*, *10*(1), 20130884.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427-430.

Rand, D. G., & Kraft-Todd, G. T. (2014). Reflection does not undermine self-interested prosociality. *Frontiers in Behavioral Neuroscience*, *8*, 300.

Rand, D. G., Newman, G. E., & Wurzbacher, O. M. (2014a). Social context and the dynamics of cooperative choice. *Journal of Behavioral Decision Making, In press*.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413-425.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014b). Social heuristics shape intuitive cooperation. *Nature Communications, 5, 3677*.

Raven, J. C. (1941). Standardization of progressive matrices, 1938. *British Journal of Medical Psychology*, *19*(1), 137-150.

Recalde, M. P., Riedl, A., & Vesterlund, L. (2014). Error prone inference from response time: The case of intuitive generosity. *CESifo Working Paper Series*, 4987.

Roca, C. P., Cuesta, J. A., & Sánchez, A. (2009). Effect of spatial structure on the evolution of cooperation. *Physical Review E*, *80*(4), 46106.

Roch, S. G., Lane, J. A., Samuelson, C. D., Allison, S. T., & Dent, J. L. (2000). Cognitive load and the equality heuristic: A two-stage model of resource overconsumption in small groups. *Organizational Behavior and Human Decision Processes*, *83*(2), 185-212.

Rubinstein, A. (2007). Instinctive and cognitive reasoning: A study of response times. *The Economic Journal*, *117*(523), 1243-1259.

Schulz, J. F., Fischbacher, U., Thöni, C., & Utikal, V. (2014). Affect and fairness: Dictator games under cognitive load. *Journal of Economic Psychology*, *41*, 77-87.

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*(3), 423-428.

Sheremeta, R., & Shields, T. (2013). Do liars believe? Beliefs and other-regarding preferences in sender-receiver games. *Journal of Economic Behavior and Organization*, *94*, 268-277.

Shinada, M., Yamagishi, T., & Ohmura, Y. (2004). False friends are worse than bitter enemies:"Altruistic" punishment of in-group members. *Evolution and Human Behavior*, *25*(6), 379-393.

Staffiero, G., Exadaktylos, F., & Espín, A. M. (2013). Accepting zero in the ultimatum game does not reflect selfish preferences. *Economics Letters*, *121*(2), 236-238.

Stanovich, K. E. (2009). *What Intelligence Tests Miss: The Psychology of Rational Thought*. Yale University Press.

Stanovich, K. E. (2010). *Rationality and the Reflective Mind*. Oxford University Press.

Stevens, J. R., & Hauser, M. D. (2004). Why be nice? Psychological constraints on the evolution of cooperation. *Trends in Cognitive Sciences*, *8*(2), 60-65.

Sutter, M., Kocher, M., & Strauß, S. (2003). Bargaining under time pressure in an experimental ultimatum game. *Economics Letters*, *81*(3), 341-347.

Sylwester, K., Herrmann, B., & Bryson, J. J. (2013). Homo homini lupus? Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics*, *6*(3), 167.

Tabibnia, G., Satpute, A. B., & Lieberman, M. D. (2008). The sunny side of fairness preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychological Science*, *19*(4), 339-347.

Tinghög, G., Andersson, D., Bonn, C., Böttiger, H., Josephson, C., Lundgren, G., et al. (2013). Intuition and cooperation reconsidered. *Nature*, *498*(7452), E1-E2.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275-1289.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147-168.

Tricomi, E., Rangel, A., Camerer, C. F., & O'Doherty, J. P. (2010). Neural evidence for inequality-averse social preferences. *Nature*, *463*(7284), 1089-1091.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, *46*(1), 35-57.

Van Lange, P. A. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology*, *77*(2), 337-349.

Van Lange, P. A., Joireman, J., Parks, C. D., & Van Dijk, E. (2013). The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*, *120*(2), 125-141.

Verkoeijen, P. P., & Bouwmeester, S. (2014). Does intuition cause cooperation? *PloS One*, *9*(5), e96654.

Williams, G. C. (1966). *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton University Press.

Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., et al. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proceedings of the National Academy of Sciences*, *109*(50), 20364-20368.

Xu, H., Bègue, L., & Bushman, B. J. (2012). Too fatigued to care: Ego depletion, guilt, and prosocial behavior. *Journal of Experimental Social Psychology*, *48*(5), 1183-1186.

Zaki, J., & Mitchell, J. P. (2011). Equitable decision making is associated with neural markers of intrinsic value. *Proceedings of the National Academy of Sciences of the USA*, *108*(49), 19761-19766.

Zaki, J., & Mitchell, J. P. (2013). Intuitive prosociality. *Current Directions in Psychological Science*, *22*(6), 466-470.

Zizzo, D. J., & Oswald, A. J. (2001). Are people willing to pay to reduce others' incomes? *Annales d'Economie et de Statistique*, *63/64*, 39-65.

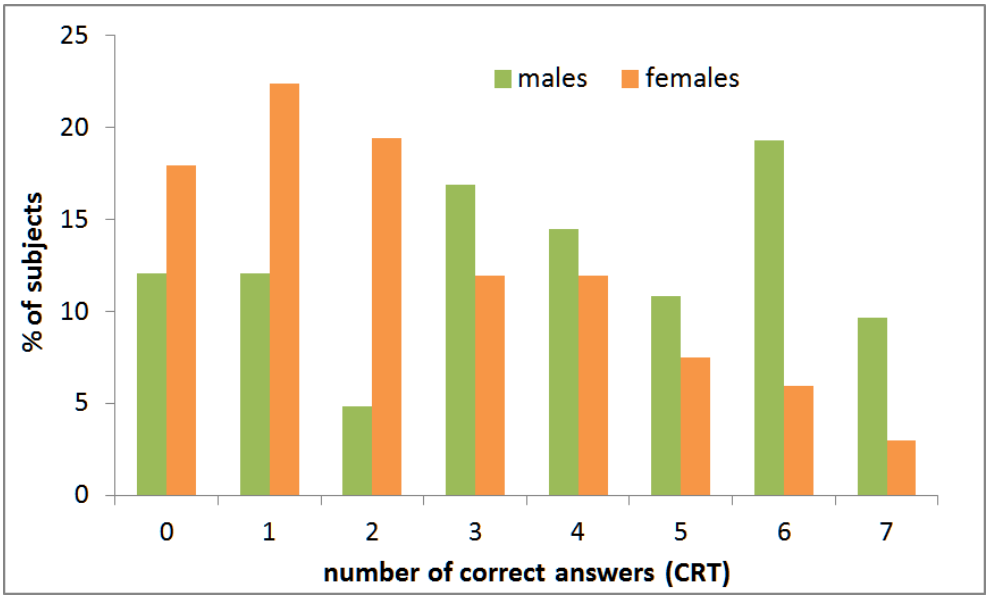

Figure S1. Distribution of CRT scores by gender (Study 1)

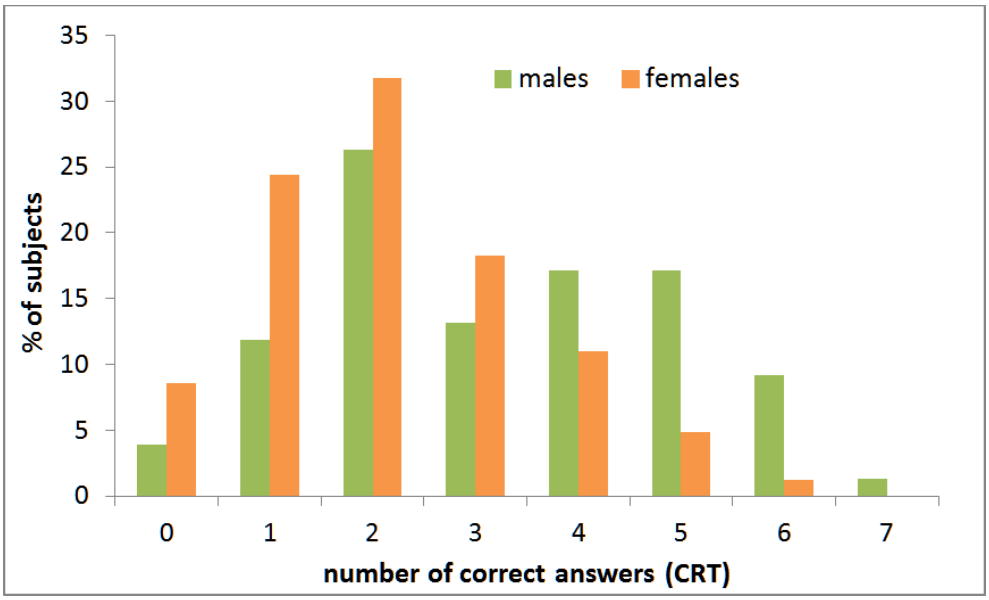

Figure S2. Distribution of CRT scores by gender (Study 2)

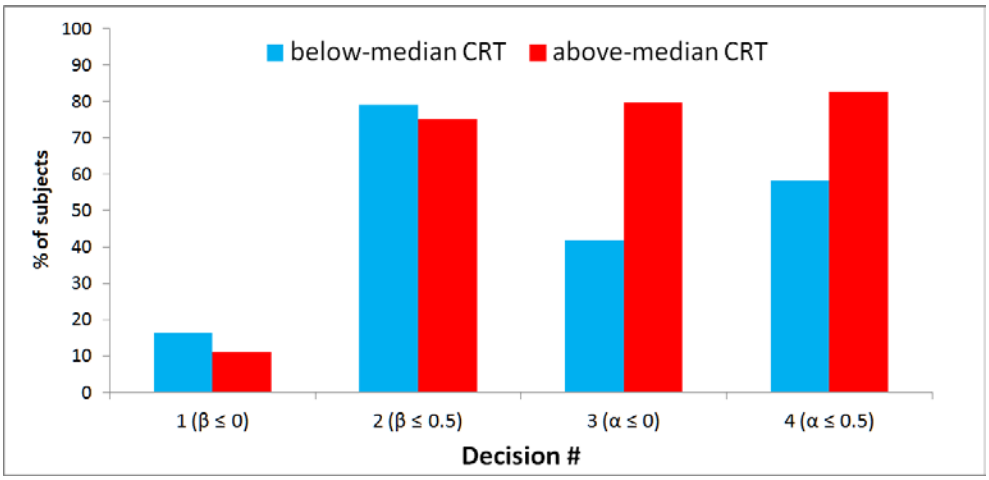

Figure S3. **Percentage of subjects choosing option B in each decision, by CRT groups (Study 1)**. The model parameters associated to option B are shown in parentheses.

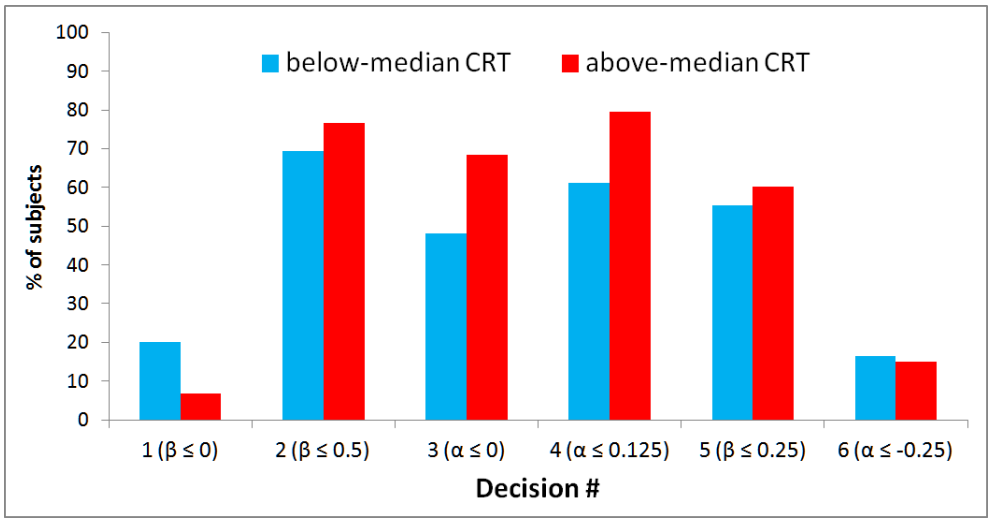

Figure S4. **Percentage of subjects choosing option B in each decision, by CRT groups (Study 2)**. The model parameters associated to option B are shown in parentheses.

Text S1. Cognitive Reflection Test

Taken from Frederick (2005):

(1) A bat and a ball cost $1.10 in total. The bat costs a dollar more than the ball. How much does the ball cost? ____ cents
[Correct answer: 5 cents; intuitive answer: 10 cents]

(2) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? ____ minutes
[Correct answer: 5 minutes; intuitive answer: 100 minutes]

(3) 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? ____ days
[Correct answer: 47 days; intuitive answer: 24 days]

Taken form Toplack et al. (2014):

(4) If John can drink one barrel of water in 6 days, and Mary can drink one barrel of water in 12 days, how long would it take them to drink one barrel of water together? _____ days
[correct answer: 4 days; intuitive answer: 9]

(5) Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are in the class? _____ students
[correct answer: 29 students; intuitive answer: 30]

(6) A man buys a pig for $60, sells it for $70, buys it back for $80, and sells it finally for $90. How much has he made? _____ dollars
[correct answer: $20; intuitive answer: $10]

(7) Simon decided to invest $8,000 in the stock market one day early in 2008. Six months after he invested, on July 17, the stocks he had purchased were down 50%. Fortunately for Simon, from July 17 to October 17, the stocks he had purchased went up 75%. At this point, Simon has: a. broken even in the stock market, b. is ahead of where he began, c. has lost money
[correct answer: c; intuitive response: b]

CRT question	Males (%)	Females (%)	p-value
1	56.63	34.33	<0.01
2	44.58	28.36	0.04
3	60.24	35.82	<0.01
4	49.40	32.84	0.05
5	42.17	19.40	<0.01
6	48.19	38.81	0.32
7	66.27	49.25	0.04

Table S1. Percentage of subjects answering correctly the CRT by question and gender (Study 1). P-values from two-sided Fisher's exact tests for the (gender) difference in proportions.

CRT question	Males (%)	Females (%)	p-value
1	47.37	31.71	0.05
2	38.16	18.29	<0.01
3	43.42	17.07	<0.01
4	47.37	24.39	<0.01
5	42.11	25.61	0.03
6	72.37	47.56	<0.01
7	31.58	53.66	<0.01

Table S2. Percentage of subjects answering correctly the CRT by question and gender (Study 2). P-values from two-sided Fisher's exact tests for the (gender) difference in proportions.

Dep var:	Decision 1 $\beta \leq 0$ (vs ≥ 0)	Decision 2 $\beta \leq 0.5$ (vs ≥ 0.5)	Decision 3 $\alpha \leq 0$ (vs ≥ 0)	Decision 4 $\alpha \leq 0.5$ (vs ≥ 0.5)
High CRT	-0.361	-0.226	0.975***	0.669***
	(0.290)	(0.236)	(0.231)	(0.238)
female	-0.512*	-0.316	-0.266	-0.313
	(0.293)	(0.235)	(0.223)	(0.228)
cons	-0.736***	0.998***	-0.059	0.383**
	(0.219)	(0.204)	(0.184)	(0.190)
ll	-58.527	-79.235	-90.056	-86.866
Wald χ^2	3.40	2.27	22.40***	12.40***
pseudo R^2	0.037	0.013	0.118	0.069

Table S3. Non-egalitarian choice (option B) as a function of CRT and Raven (Study 1). Probit estimates. High CRT is a dummy that takes value 1 if the CRT score is above the median and takes value 0 otherwise. Robust standard errors clustered on individuals are shown in parentheses. *, **, *** denote significance at the 0.10, 0.05 and 0.01 level, respectively. $N = 150$ in all regressions.

Dep var:	Decision 1 $\beta \leq 0$ (vs ≥ 0)	Decision 2 $\beta \leq 0.5$ (vs ≥ 0.5)	Decision 3 $\alpha \leq 0$ (vs ≥ 0)	Decision 4 $\alpha \leq 0.125$ (vs ≥ 0.125)	Decision 5 $\beta \leq 0.25$ (vs ≥ 0.25)	Decision 6 $\alpha \leq -0.25$ (vs ≥ -0.25)
High CRT	-0.585**	0.221	0.460**	0.417*	0.178	-0.082
	(0.281)	(0.222)	(0.209)	(0.217)	(0.208)	(0.254)
female	0.318	-0.003	-0.310	-0.504**	0.220	-0.118
	(0.270)	(0.221)	(0.209)	(0.216)	(0.208)	(0.252)
cons	-1.049***	0.510**	0.149	0.614***	-0.004	-0.905***
	(0.241)	(0.199)	(0.189)	(0.189)	(0.188)	(0.229)
ll	-60.046	-91.959	-103.249	-91.269	-106.920	-68.856
Wald χ^2	7.11**	4.05	8.49**	9.48***	1.52	0.25
pseudo R^2	0.058	0.006	0.041	0.059	0.007	0.002

Table S4. Non-egalitarian choice (option B) as a function of CRT (Study 2). Probit estimates. High CRT is a dummy that takes value 1 if the CRT score is above the median and takes value 0 otherwise. Robust standard errors clustered on individuals are shown in parentheses. *, **, *** denote significance at the 0.10, 0.05 and 0.01 level, respectively. $N = 158$ in all regressions.