Cognitive Reflection Test: Whom, How, When

Pablo Brañas-Garza[1,2], Praveen Kujal[3], Balint Lenkei[3]

The use of the Cognitive Reflection Test as a covariate to explain behavior in Economics and Psychology experiments has significantly increased in the past few years. Experiments have shown its usefulness in predicting behavior. However, little is known about if the test is gender biased, whether incentives matter or how different implementation procedures impact outcomes. Here we report the results of a meta-study of 118 Cognitive Reflection Test studies comprising of 44,558 participants across 21 countries. We find that there is a negative correlation between being female and the overall, and individual, correct answers to CRT questions. Monetary incentives do not impact performance. Regarding implementation procedures, taking the test at the end of the experiment negatively impacts performance. Students perform better compared to non-students. We obtain mixed evidence on whether the sequence of questions matters. Finally, we find that computerized tests marginally improve results.

Keywords: CRT, Experiments, Gender, Incentives.

JEL Classification: C90, C91, C93.

[1] Loyola Behavioral Lab, Universidad Loyola Andalucia, Escritor Castilla Aguayo, 4, 14004 Córdoba, Spain

[2] Corresponding author: branasgarza@gmail.com.

[3] Economics Department, Middlesex University London, The Burroughs, London, NW4 4BT, U.K.

1. Introduction

In this meta-study we test for several of the empirical regularities regarding the Cognitive Reflection Test (Frederick, 2005) reported in several Economics and Psychology experiments. We have a heterogenous sample of studies characterized by differences in geographical location, incentives, non-student samples, lab/field based, etc. We test for whether the reported gender differences hold and whether monetary incentives significantly impact the number of correct responses in the Cognitive Reflection Test (henceforth CRT). Our meta-study also compares the CRT results for student and non-student samples of participants. We also test for whether different procedures such as the timing of the CRT, the use of computerized settings, or increased exposure to the CRT over the years has any impact on the observed results.

The CRT was first proposed by Frederick (2005) and since then has been extensively used in the Experimental Economics and Psychology literature. Frederick proposed the test based on a dual-system theory (e.g. Epstein 1994; Sloman 1996; Stanovich and West 2000; Kahneman and Frederick 2002) made up of two cognitive processes: System 1, executed quickly without much reflection and System 2, more deliberate and requiring conscious thought and effort. The questions in the CRT have an immediate (intuitive) incorrect response (System 1). However, the correct response requires some deliberation, i.e. the activation of System 2. The standard CRT test consists of the following three questions:

- *A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? (Intuitive answer 10, correct answer 5).*
- *If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? (Intuitive answer 100, correct answer 5).*
- *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? (Intuitive answer 24, correct answer 47).*[4]

[4] We will refer to the first, second and third questions as "B&B" (Bat and Ball), "Machines" and " Lillypad", respectively.

Frederick (2005) found that individuals with high CRT scores are more patient and more willing to gamble in the domain of gains. He also provided evidence that the CRT scores are highly correlated with some other tests of analytic thinking (e.g. ACT, SAT and WPT) and that males on average score higher on the test. Toplak et al. (2011) claim that the CRT can be viewed as a combination of cognitive capacity, disposition for judgement and decision making. They argue that the CRT captures important characteristics of rational thinking that are not measured in other intelligence tests. Below we discuss the results from CRT related studies.

Since Frederick (2005), several researchers have adopted the CRT as a measure of cognitive abilities and have studied its predictive power in decision making (e.g. Oechssler et al. 2009; Campitelli and Labollita 2010; Hoppe and Kusterer 2011; Besedes et al. 2012; Andersson et al. 2013; Moritz et al. 2013 Neyse et al. 2016). Oechssler et al. (2009) investigate whether behavioral biases are related to cognitive abilities. Replicating Frederick (2005), they find that participants with low scores on the CRT are more likely to be subject to the conjunction fallacy and to conservatism in updating probabilities (also see Liberali et al. 2012; Alós-Ferrer and Hügelschäfer 2016). Meanwhile, Bosch-Doménech et al. (2014) find biological underpinning´s for CRT performance relating the 2D:4D ratio and performance on the CRT. They find that a lower 2D:4D ratio (reflecting a relative higher exposure to testosterone) is significantly associated with higher scores on the CRT. Neyse et al. (2016) find that controlling for cognitive capacity using CRT they find that subjects with higher financial literacy are better in choosing optimal investments.

Some authors have also looked at the role of cognition in experimental consumer goods markets (Alexander et al. 2018) where System 1 responses in economic transactions can alter market equilibrium. Buyers may intuitively feel that the correct response to a high-pressure offer is to reject it, thereby reducing the profitability of that technique. Others have found that high CRT scores are significantly related to search efforts (Lukas et al., 2019) in experimental markets. Königsheim, Lukas, Nöth (2019) have meanwhile used the CRT to study how subjects focus on the task at hand. Sheremeta, R. M. (2018) has finds that participants who are more impulsive overbid more in contests. Meanwhile, Ruffle and Wilson (2018) use CRT to study risk behaviour is tattooed individuals.

The CRT has also been found to be a good predictor of the degree of strategic behavior in laboratory experiments (e.g. Brañas-Garza et al. 2012; Carpenter et al. 2013; Kiss et al. 2016 etc.). It is a useful test to measure strategic behavior as it not only captures reflective processes but also the ability to execute small computational tasks (Corgnet et al. 2015). Brañas-Garza et al. (2012) investigate the relationship between CRT outcomes and subject performance in the repeated feedback-free p-Beauty Contest Game (Nagel 1995), where a higher level of reasoning indicates better strategic behavior. They find that individuals with higher scores on the CRT choose numbers closer to the Nash equilibrium meanwhile they find that other measures of cognitive abilities (e.g. Raven) have a smaller or not significant effect in predicting subjects' choices. Kiss et al. (2016) look at the effect of CRT on withdrawal decisions in an extended version of Diamond and Dybvig's (1983) bank-run game. They find that participants with higher cognitive abilities (as measured by the CRT) identify the dominant strategy when strategic uncertainty is present in the game. The above evidence indicates that the CRT could also help us in identifying strategically sophisticated subjects[5].

It is now well established in the Behavioural Economics and Psychological literature that subjects with greater cognitive abilities are other-regarding (e.g. Ben-Ner et al. 2004; Chen et al. 2013). In recent years the link between CRT scores and social preferences has been investigated (Corgnet et al. 2015; Peysakhovic and Rand 2015; Ponti and Rodriguez-Lara 2015; Cueva-Herrero et al 2016). Corgnet et al. (2015) find that individuals with a high CRT score are more likely to make altruistic choices in simple non-strategic decisions. Their choices increase social welfare by increasing the other person's payoff at a very low (or none) cost for the individual. On the other hand, the choices of less reflective subjects are more correlated with spiteful motives.

There is also evidence regarding the relationship between behavioral biases and cognitive reflection in the literature on behavioral finance and experimental asset markets (e.g. Cheung et al. 2014; Noussair et al. 2014; Corgnet et al. 2014; Bosch-Rosa et al. 2015; Holt et al. 2015 etc.). Corgnet et al. (2014) find that high CRT subjects earned significantly more on average than the initial value of their portfolio while low CRT subjects earned less. Interestingly, subjects with low CRT scores were net purchasers (sellers) of shares when the price was above (below) fundamental value while the opposite was true for subjects with high CRT scores. Bosch-Rosa et al. (2015) run a battery of tests to assess subjects cognitive sophistication and then classify subjects into low or high levels. They find that if subjects with only low cognitive abilities are trading in an experimental asset market it will lead to the classic asset

[5] An exception to this Hanaki et al (2016) who find that Raven's test score is a more reliable predictor of strategic behavior than CRT score: whenever the latter predicts behavior, the former does too, but not vice versa. Subjects with higher Raven's test scores are more likely to use the dominant strategy and to best respond to other player's dominant strategy. Unlike those with low Raven's test score, they also react to the presence of strategic uncertainty.

market bubble. However, asset markets with only high cognitive sophistication trade close to their fundamental values. In a recent paper Holt et al. (2015) study gender differences in an experimental asset market where participants answer the standard CRT questions (with an additional mathematical question). Though they observe no gender differences in bubble formation, they find that male subjects performed better on all questions, and the difference was largest for the more mathematical (speed) question.

Another important issue is regarding gender differences. It has been shown that males consistently score significantly higher on the CRT than females (e.g. Frederick 2005; Hoppe and Kusterer 2011; Holt et al. 2015; Cueva-Herrero et al. 2016 etc.). This agrees with the findings in the experimental literature that show that males have higher mathematical abilities and score higher than females on math tests (e.g. Benbow and Stanley 1980; Aiken 1986-1987; Benbow et al. 2000; Mau and Lynn 2010 etc.).

An important question in both Economics and Psychology is regarding the use of incentives in experiments. The effect of incentives on CRT responses has not been directly studied so far. The available evidence regarding how incentives affect outcomes is split, i.e. whether incentives matter or not is context dependent. For example, Riedel et al. (1988), Scott et al. (1988) and Duckworth et al. (2011) find a positive relationship between monetary incentives and performance levels meanwhile, others (e.g. Jenkins et al. 1998; Camerer and Hogarth 1999; Bonner and Sprinkle 2002) find evidence to the contrary. Studies that reject the impact of monetary incentives on performance outcomes argue that while it increases effort, it either doesn't improve performance at all or it only increases the performance of those who possess better cognitive abilities (Awasthi and Pratt 1990)[6].

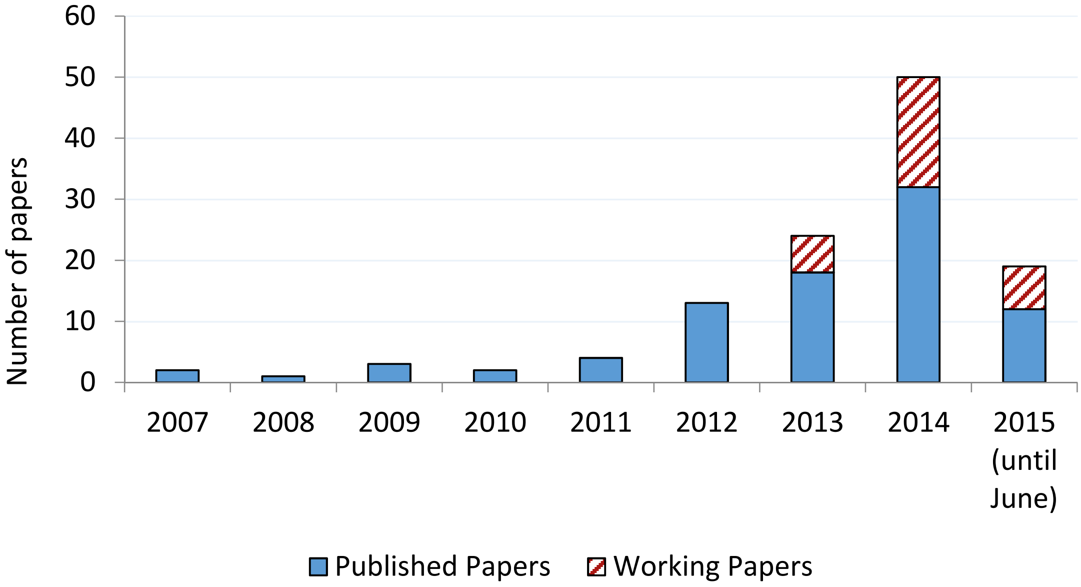
Another issue has to do with external validity of laboratory experiments. That is, it is not clear as to how much the results from the laboratory (with university students) can be extrapolated to choices made by non-students. The evidence, again, is mixed. That is, there are mixed views on whether studies conducted with (volunteering) university students provide reliable results (Peterson 2001; Levitt and List 2007; Falk and Heckman 2009; Falk et al. 2013; Exadaktylos et al. 2013). Common objections are that student subject pool sample sizes are small and not representative. Given our large sample we can address this issue in the context of the CRT.

[6] The cognitive characteristic examined by Awasthi and Pratt (1990) is perceptual differentiation (PD) i.e. an individual's ability to perceptually abstract from a complex setting certain familiar concepts or relationships.

We also look at the effect of positioning of the CRT compared to the main experiment. If CRT is of interest in finding covariation with decisions made in the experiment then it becomes important to understand whether timing matters. If cognitive load diminishes decision making ability then we would expect better performance the earlier is the test taken in the experiment.

Finally, we study the issue of prior experience with the CRT. This has to do with the point made by Toplak et al. (2014) where they argue that if the CRT is commonly used it is probable that individuals may have become familiarized with it. Figure 1 presents the total number of working and published papers included in our analysis over the period of 2007 to 2015. It is clear that in recent years the CRT has been increasingly used. The direct concern raised by Toplak et al (2014) is difficult to test with our data as we have no information on repeat takers of CRT. Further, note that this matter is confounded with the frequent use of the Amazon Mechanical Turk (AMT henceforth) platform for running experiments (for review on AMT see Paolacci et al. 2010; Buhrmester et al. 2011; Goodman et al. 2013). Given this study whether the year a particular CRT study was conducted and whether it was conducted on line affected test scores. This issue is also related whether different administration modes, i.e. computerized or paper-and-pencil, provide significantly different outcomes (e.g. George et al. 1992; King and Miles 1995; Cole et al. 2006 etc.). We will also be studying this with our data.

Figure 1. Number of papers in our meta-study according to the year the papers were published



The paper is organized as follows. Section 2 presents the procedures and techniques used for data collection. Section 3 provides the results. Section 4 concludes. All additional information is in the Appendix.

2. Procedures

2.1. Data collection

The information and data on the CRT were obtained through two channels. First, an e-mail inviting members of the Economic Science Association (ESA) was sent. In addition, a reminder e-mail was sent before the process was closed in June 2015. Respondents were provided with an online survey where they could input information about their study. Figure A1 (Appendix 2) presents a screen shot of the actual questionnaire that researchers were asked to fill out.

Second, we searched for research articles using the phrase “Cognitive Reflection Test” on Google Scholar. If an article was identified as one where the CRT was conducted the corresponding author was e-mailed the survey. The researchers were asked to respond to the following questions on the survey:

- Total Number of CRT participants (and the number of females among the total).
- How many of the total answered the *B&B*, *Machines*, and *Lillypad* questions correctly (and the number of females among them).
- Out of the total how many participants answered all *Three*, *Two* or *One* question(s) correctly (and the number of females among them).
- Whether the subjects received monetary incentives for correct answers.
- Whether the CRT was computerized or it was a paper and pencil test.
- The order of the CRT questions.
- Whether the CRT was conducted before, in-between or after the experiment.

We contacted 190 authors through e-mail and received information on 118 studies (62%) in total through filling out the survey (in some cases the authors had multiple studies). The corresponding authors we contacted (based on our Google Scholar search) represent roughly all the papers we could identify that used the CRT. Due to a considerable number of invitees declining to participate, our study may be hampered by self-selection bias. However, some degree of self-selection when inviting researchers to participate in a meta-analysis is almost inevitable. We still managed to obtain studies from a wide range of disciplines, both published and unpublished, and have considerable heterogeneity in our data.[7]

2.2. Sample creation

Appendix B provides a list of all research articles included in our analysis. Some research papers in our meta-analysis include two or more CRT studies. Overall our data comprises of 118 studies with 44,558 participants between the years 2007 and 2015. The articles represent a wide range of disciplines including Economics (58.1% of studies), Psychology (33.3%) and Management (2.8%) with researchers from 21 different countries[8]. The largest number of studies was conducted in the USA and Germany, 42 and 15, respectively. The study with the lowest number of observations was 40, while the study with the most had 4,312. Table 1 includes a breakdown regarding the number of observations available in each category in our sample.

The full sample of 44,558 subjects was broken down into further sub-categories. These were:

- *Female* (vs Male=0).
- Monetary *incentives* (whether the experimenter paid monetary incentives for correct answers).
- *Students* (vs Non-students=0).
- Position (whether the CRT was conducted *before*, *in-between* or *after* experiments).
- *Visibility* (the year in which the studies were conducted, see also Table 4).
- *Sequence* (the order in which the CRT questions were asked).
- *Computerized* (vs paper and pencil=0).

[7] However, we did not ask for papers for a specific purpose (e.g. link between the CRT and playing the Nash Equilibrium). In this sense we may expect less self-selection.

[8] These countries include (in alphabetical order): Argentina, Australia, Austria, Brazil, Canada, China, Colombia, Denmark, Finland, France, Germany, Israel, Italy, Japan, Netherlands, Slovakia, Spain, Sweden, Switzerland, UK and USA.

Table 1. Data Distribution

	Distribution (full sample)	**Distribution (regression)**
Number of studies	**118**	**118**
Total number of observations	**44,558**	**39,603**
N (Bat and Ball, Machines, Lillypad correct answers)	**41,004**	**38,031**
Bat and Ball correct	31.75%	32.24%
Machines correct	40.24%	40.84%
Lillypad correct	47.78%	48.59%
N (3,2,1 and None correct answers)	**44,558**	**39,603**
All 3 answers correct	18.17%	18.64%
Only 2 answers correct	21.12%	21.45%
Only 1 answers correct	23.18%	23.33%
None of the answers correct	37.54%	36.57%
N (gender)	**41,705**	**39,603**
Female	52.76%	52.89%
Male	47.24%	47.11%
N (monetary incentives)	**44,558**	**39,603**
Incentivized	14.67%	15.82%
Non-Incentivized	85.33%	84.18%
N (student)	**43,684**	**39,603**
Student	42.28%	41.42%
Non-Student	57.72%	58.58%
N (position of the test)	**44,558**	**39,603**
CRT took place before the experiment	37.66%	34.77%
CRT took place after the experiment	44.58%	46.46%
CRT took place in-between experiments	17.75%	18.77%
N (sequence of the questions)	**44,558**	**39,603**
Questions asked in standard sequence (B&B, Machines, Lillypad)	83.78%	84.92%
Questions asked in randomized sequence	11.64%	13.09%
Questions asked in B&B; Lilly Pad; Machines sequence	0.90%	1.01%
Questions asked in Machines; Lilly Pad; B&B sequence	2.82%	0%
Questions asked in Lilly Pad; B&B; Machines sequence	0.87%	0.97%
N (computerized or paper and pencil)	**42,797**	**39,603**
Computerized	87.91%	89.65%
Paper and Pencil	12.09%	10.35%
N (country information)	**44,217**	**39,603**
Anglo-Saxon	49.65%	46.59%
Europe	41.65%	43.70%
Rest of the world	8.70%	9.71%

2.3. Empirical strategy

We use OLS regressions to estimate the relationship between CRT outcomes and the list of variables defined earlier.[9] We use the OLS as the interpretation of its coefficients is direct. The robust standard errors are clustered around study ID's. Our meta-analysis includes 118 studies with substantial heterogeneity (e.g. paper and pencil/computerized; incentivized/non-incentivized etc.). In order to check for the robustness of our analysis we re-run our main regressions (Table 3) with six additional sub-samples (see Appendix):

- A sub-sample including female subjects only (Appendix Table ER1). In section 3.1 we analyze the impact of gender differences on CRT results.
- A sub-sample excluding studies where monetary incentives were used to reward correct answers (Appendix Table ER2). In section 3.2 we analyze the impact of monetary incentives on CRT performance.
- A sub-sample comprised of non-students (Appendix Table ER3). In section 3.3 we analyze the difference in CRT results between university student samples and samples including non-students.
- A sub-sample excluding the studies where experiments were not conducted (Appendix Table ER4). In section 3.4 we analyze the impact of positioning of the CRT compared to the main experiment (i.e. before, in-between or after). Our general sample includes studies where the researchers did not run experiments. Having these observations in our sample could potentially lead to biased estimates. Further, by excluding these observations we can isolate the effect of these studies on the positioning of the CRT test.
- A sub-sample excluding studies where the experimenters used Amazon Mechanical Turk (Appendix Table ER5). In section 3.5 we discuss subjects' exposure to the CRT over the years. Popular online experimental platforms such as the AMT may have made the test more visible over the years. Further, the ease of access to the correct answers raises important methodological concerns[10].

[9] Other statistical models such as probit and logit provide similar results (see Appendix).

[10] We instantly obtained answers to all three questions through Google search.

- A sub-sample excluding the studies where the sequence of the questions were randomly determined (Appendix Table ER6). In section 3.6 we analyze the effect of the CRT question sequences on test outcomes. We divide our full sample between standard sequence (i.e. *B&B, Machines, Lillypad*) and other sequences. The general sample however includes studies where the sequence of questions is randomly determined. There is a 1 in 6 chance that randomization generates a standard sequence. By excluding random sequences we can isolate the effect of having standardized sequences in the other sequence sub-sample.

Table 2. Mean test scores

	B&B	***Machines***	***Lillypad***	***None***	***1***	***2***	***3***
Gender							
Male	38.37%	50.43%	59.02%	27.01%	22.78%	25.00%	25.21%
Female	26.70%	32.18%	39.18%	45.09%	23.83%	18.29%	12.79%
Monetary incentives							
No monetary incentives	31.74%	39.50%	47.14%	37.82%	23.30%	20.83%	18.06%
Monetary incentives	34.76%	47.60%	55.95%	29.96%	23.53%	24.77%	21.74%
Non students vs. Students							
Non-Student	26.07%	39.21%	45.61%	40.60%	23.23%	20.05%	16.12%
Student	40.68%	43.06%	52.68%	30.88%	23.48%	23.44%	22.21%
CRT positioning							
Before	31.34%	41.94%	53.71%	33.70%	24.48%	23.04%	18.78%
In-between	32.54%	38.38%	47.30%	37.73%	23.84%	20.92%	17.51%
After	32.73%	41.08%	45.56%	38.25%	22.28%	20.48%	18.99%
Question ordering							
Non-standard order	22.60%	33.07%	34.30%	51.02%	20.31%	15.82%	12.84%
Standard order	34.01%	42.27%	51.23%	34.01%	23.87%	22.45%	19.67%
Paper and Pencil vs. Computerized							
Paper and Pencil	37.14%	36.94%	42.86%	38.66%	22.78%	19.73%	18.83%
Computerized	31.67%	41.28%	49.25%	36.33%	23.40%	21.65%	18.62%

Notes: The first three columns refer to N= 38031, while the other four columns refer to N= 39603

3. Results

We now look at how the questions were answered both individually and overall. Figure 2 shows a summary of the results for the correct answers (by question) and for the entire test. The left side refers to the number of correct answers for each question, i.e. *B&B, Machines and Lillypad* (N = 41,004). While the *B&B* question was answered correctly by 32% in the sample, the fraction rises to 48% for the *Lillypad* question. It is hard to interpret what these proportions mean. Either the *B&B* question is more cognitively demanding for the subjects, or non-incentivized implementation (or cognitive laziness) may imply that subjects only answered the "more" intuitive questions first and did not bother answering the more cognitively difficult[11]. The two-tailed t-tests (equal/unequal variances) comparing the means of the *B&B, Machines, Lillypad* distributions reject the null hypothesis of equal means ($p<0.01$).

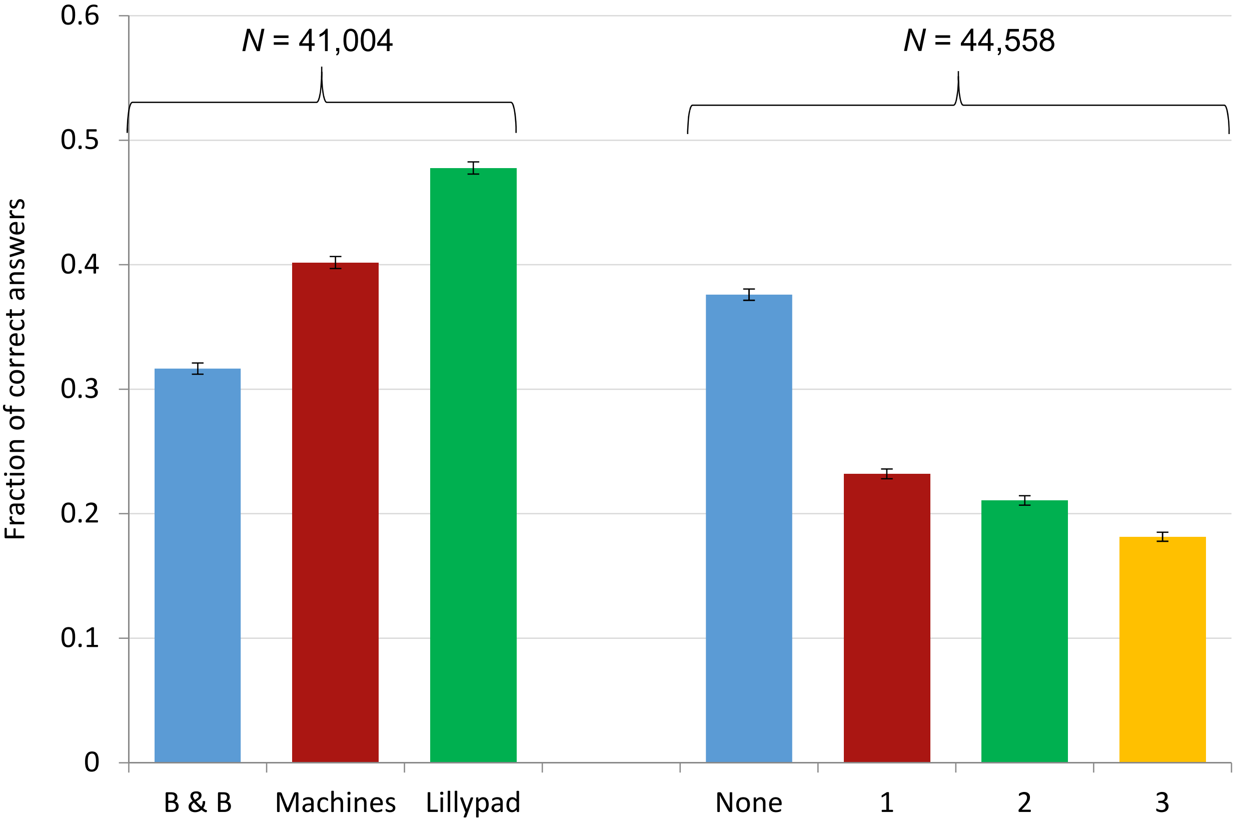
Looking at the total number of correct answers (right hand side, Figure 2)[12] we find that a third of the population lack reflective, or cognitive, abilities. Meanwhile, the remaining 62% have at least some, including 18% that provide all correct answers. Two-tailed t-tests (equal/unequal equal variances) comparing the distribution of the *None, 1, 2, 3* correct answers reject the null hypothesis of equal means everywhere ($p<0.01$).

Next, we study in detail the determinants of correct answers to the CRT. Moreover, in order to better understand these estimates we run a series of additional regressions in Appendix 1. Specifically, we repeat the main regression using a subsample of females only (Table ER1), a subsample of studies without monetary incentives (Table ER2), a subsample of non-students (Table ER3), a subsample of studies without economic experiments (Table ER4), a subsample excluding Amazon Mechanical Turk studies (ATM, Table ER5) and lastly a subsample of study excluding CRTs with random order (Table ER6).

[11] Note that under lack of incentives participants may choose not to answer the cognitively difficult question.

[12] Note that differences in the sample sizes are due to data availability.

Figure 2. The fraction of correct answers in the meta-study.



3.1. Gender bias

Table 2 shows that the CRT has a strong gender bias favoring males (*N* = 41,705; females 52.76%). We find that: (i) males perform better in every single question, (ii) females are more likely to answer none of the questions correctly, and (iii) males are more likely to answer all three questions correctly.

Importantly, gender differences persist in a regression (Table 3, row 1) even when we control for test characteristics (e.g. monetary incentives, computerized, student samples, positioning of the experiment etc.). Our results confirm Frederick (2005) (*N* = 3,428) who showed that males perform better in the CRT (also see Oechssler et al. 2009; Hoppe and Kusterer 2011; Holt et al. 2015; Cueva-Herrero et al. 2016 etc.).

We replicated the regressions in studies without incentives (Tables ER2), in a subsample of non-students (ER3), in studies without experiments (ER4), in Non-AMT studies (ER5), and in studies without randomly sorted questions (ER6). Our replications show that the gender bias remains negative and statistically significant (p<0.01) throughout. Hence, we find that all previous results hold.

In sum, gender has an important impact on CRT performance and if used as a sorting criteria may bias the distribution of participants. This gives us our first result,

Result 1: Our results strongly support that CRT responses have a strong gender bias. The proportion of males is increasing with increase in the score.

This is a useful result as knowing that the CRT has a strong male bias can be important for sample building. For instance, say that we would like to select subjects with certain characteristics from the sample. Our study suggests that using the 3-correct-answers criteria will give us twice as many males than females. This implies that we not only select highly cognitive individuals, but also that the sample is strongly biased towards males.

Bosch-Rosa et al. (2015), for example, divide their subject pool between individuals with low and high cognitive abilities based on the CRT results in order to perform a later task. Our results suggest that their findings might be partly driven by gender effects. A similar problem arises in Brañas-Garza et al. (2012) where they find that high CRT scorers are more likely to play according to the Nash Equilibrium in the Beauty Contest Game. This may again be due to the higher proportion of males rather than just an overall effect of high CRT scorers.

Finally, some recent authors (Sirota et al. (2018) and Thomson and Oppenheimer (2016) have argued that the CRT is gender biased due to its mathematical component and familiar. They argue that the traditional CRT confounds reflection ability with mathematical ability. Sirota et al (2018) develop a verbal CRT to address this issue and find that their test did not have a gender bias. Regarding the question of increased familiarity with the test it is worth pointing out a recent study by Stagnaro, Pennycook and Rand (2018) has shown that performance across the three question CRT is stable across time.

Table 3. Regression analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	B&B	*Machines*	*Lillypad*	*None*	*1*	*2*	*3*
(1) female	-0.113***	-0.177***	-0.197***	0.179***	0.009	-0.066***	-0.121***
	(0.011)	(0.010)	(0.010)	(0.010)	(0.006)	(0.007)	(0.008)
(2) monetary incentives	-0.026	0.003	0.040	-0.005	-0.002	0.000	0.008
	(0.046)	(0.048)	(0.049)	(0.045)	(0.016)	(0.017)	(0.040)
(3) student	0.138***	-0.002	0.067*	-0.089**	0.011	0.030**	0.047*
	(0.035)	(0.025)	(0.039)	(0.034)	(0.008)	(0.013)	(0.024)
(4a) in-between experiments	-0.046	-0.007	-0.090*	0.059	0.002	-0.017	-0.043
	(0.045)	(0.035)	(0.049)	(0.040)	(0.013)	(0.014)	(0.033)
(4b) after the experiment	-0.032	-0.009	-0.093**	0.060*	-0.008	-0.026**	-0.026
	(0.037)	(0.030)	(0.038)	(0.035)	(0.009)	(0.012)	(0.026)
(5) visibility	0.008	0.016***	0.005	-0.005	-0.007***	0.002	0.010**
	(0.006)	(0.006)	(0.006)	(0.005)	(0.002)	(0.002)	(0.005)
(6) standard sequence	0.103**	0.102***	0.148***	-0.142***	0.012	0.050***	0.080***
	(0.040)	(0.034)	(0.043)	(0.040)	(0.012)	(0.015)	(0.031)
(7) computerized	0.033	0.085*	0.108**	-0.095**	0.013	0.050**	0.032
	(0.038)	(0.048)	(0.051)	(0.045)	(0.012)	(0.020)	(0.032)
constant	0.184**	0.270***	0.285***	0.533***	0.241***	0.156***	0.070
	(0.072)	(0.073)	(0.074)	(0.074)	(0.022)	(0.030)	(0.056)
N	38031	38031	38031	39603	39603	39603	39603
R^2	0.045	0.052	0.071	0.067	0.003	0.015	0.038

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, *p<0.1. The regressions also control for country by using two dummies: europe and anglo-saxon.

3.2. Incentives

The effect of financial incentives on human behavior has been a long debated issue in Economics and Psychology (for a review see Camerer and Hogarth 1999). The dominant argument in the experimental methodology is that incentives are important for profit maximizing individuals. In our case this would imply that the number of correct answers would improve under monetary incentives (14.67% of our sample).

The regression analysis (row 2, Table 3) shows that the variable *monetary incentives* is not statistically significant at any of the common significance levels. This implies that paying subject for correct answers on the CRT does not increase performance levels.

Our robustness checks show that effect of incentives are only marginally significant for the female subsample (Tables ER1), for non-students (ER3) and in studies without experiments (ER4); while the lack of-effect of *monetary incentives* remains persistent throughout in Non-AMT studies (ER5) and in studies without randomly sorted questions (ER6). Below we present our result,

Result 2: Our results support that overall incentives have no impact on CRT performance.

Note that the role of incentives and the degree of cognition can also be important. For example, Awasthi and Pratt (1990) find that the effectiveness of monetary incentives depends on the cognitive skill of the decision maker. In their study monetary incentives were associated with higher performance only for higher cognition individuals. We cannot comment on whether there is a relation between cognition and incentives. One may also argue that the test was a marginal part of a larger study and payments were not salient (Gneezy and Rustichini 2000). Finally, we should point out that we lack specific details on how incentives were implemented and their magnitude.

Finally, it is important to point out that Neyse et al. (2016) find that both CRT scores and expectations (regarding the score) are higher in a modified CRT with incentives. Average correct answers to the CRT were 4.04 without incentives and 4.46 with incentives (rank-sum p = 0.033). Meanwhile, expected own performance for correct answers increased from 5.18 to 5.74 (rank-sum p = 0.000).

3.3. Students vs. non-students

Economics experiments have been traditionally run with university students. This has raised an obvious question about external validity of experimental data (see Levitt and List 2007; Falk and Heckman

2009; Exadaktylos et al. 2013). Our sample includes studies that were conducted with, and without, university students (42.28% of all observations). This allows us to check for whether there are population differences in the CRT.

Overall we find that the student population performs better than non-students. We find that students score significantly better in the *B&B* and, only slightly better in the *Machines* and *Lillypad* question (Table 2). The results in Table 2 also show that university students are less likely to have all three questions answered incorrectly, while at the same time they are more likely to give two and three correct answers. Below we summarize our results,

Table 3 (row 3) confirms the findings in Table 2. The *student* coefficient is statistically significant for the *B&B (p<0.01)* and *Lillypad (p<0.1)* questions implying that students are more likely to give correct answers to these two questions. In contrast, the coefficient for zero correct answers is negative and statistically significant at the *5%* level. This implies that non-students on average are more likely to obtain all incorrect answers relative to students. Furthermore, students are more likely to have two *(p<0.05)* and all three *(p<0.1)* answers given correctly. Results on the high performance of students compared to non-students are likely to be derived from higher cognitive ability of students compared to average population (e.g. Pennycook et al. 2012).

Our robustness checks show that these effects have similar signs but less statistical power for the female subsample (Table ER1), subsample without monetary incentives (ER3), in studies without experiments (ER4) and in studies without randomly sorted questions (ER6), however similar significance levels for the subsample using only Non-AMT studies (ER5). In sum, our results allow us to state that one can expect the average CRT scores to be higher when using *student* samples.

Result 3: We find that students perform significantly better than non-students.

3.4 When?

If CRT is of interest in finding covariation with decisions made in the experiment then it becomes important to understand whether timing matters. In our sample the proportion of studies where the test was conducted *before*, *in-between* or *after* the experiment is 37.66%, 17.75% and 44.58%, respectively. A priori one would expect no differences. However, there are reasons why the timing may be important. The first is cognitive load. If students perform cognitively difficult tasks in the experiments then a later CRT would imply higher cognitive load and hence may affect CRT response rates. The second argument could be related to glucose depletion. It has been shown that brain activity is reliant on blood glucose levels as it affects the firing of neurons (Weiss 1986). Experimental tasks almost always require some form of cognition (reading instructions, answering questionnaires, quizzes etc.) and it would be reasonable to assume that glucose levels would be lower towards the end of the experiment. This would then consequently imply that if the CRT is conducted at the end of the experimental then performance on the CRT should be negatively affected[13]. Below we present our results,

The main message from our analysis is that CRT efficiency declines the later it is conducted. One sees that there are some differences in CRT performance depending upon whether it was conducted *before*, *in-between* or *after* the experiment (see rows 4a and 4b in Table 3). Conducting it *in-between* or *after* has a negative and statistically significant effect on the *Lillypad* question (*$p<0.1$* and *$p<0.05$*, respectively) (rows 4a and 4b, Table 3). In addition, conducting it *after* is more likely to result in *None (p<0.1)* and less likely to have exactly two questions answered correctly *(p<0.05)*. It is important to note that the *after-the-experiment* coefficient remains negative throughout (row 4b, Table 3). This suggests that conducting the CRTs *after the experiments* can potentially impact outcomes negatively.

Note, however, that prior data includes studies where no experiments were conducted. We conducted further analysis by removing these studies from the sample. This gives us even stronger results (Table ER4, rows 4a and 4b). Now it is even less likely that subjects are to answer the *B&B* and *Lillypad* questions correctly if CRTs conducted *in-between* or *after the experiments*. This negative effect is lower for *in-between* experiments *(p<0.05)* and stronger for *after* the experiments *(p<0.01)* variables. The stronger negative effect for the variable *after* is coherent with the argument that glucose levels are being depleted as subjects are progressing through the experiment. Similarly, we observe that subjects are less likely to answer all three questions correctly both *in-between* and *after experiments* (both *$p<0.05$*) and more likely to have *None* (both *$p<0.05$*) (rows 4a and 4b, Table ER4).[14]

Result 4: Performance in the CRT improves the earlier the test administered in the experiment. The results are stronger excluding CRT implementation without an experiment.

[13] People performing worse on the CRT at the end of the experiment can be also confounded by the fact that they may have had less time then. For example, the experiment running late. In addition, experiments that measure CRT at the end may be different and typically longer.

[14] The robustness checks (Tables A3, A4, A6, A8) report similar results with varying degrees of significance.

Whether it is cognitive load or glucose depletion it is important to know that performance in the test gets worse the later it is conducted in the experiment. It is known that that glucose levels (in the brain) play an important role in cognition. Effortful, controlled or executive processes and tasks (e.g. experiments) require more glucose than simpler, less effortful or automatic processes. When glucose levels are low, cerebral functioning is disrupted, producing numerous cognitive and behavioral deficits (Gailliot and Baumeister 2007). In sum, our results show that conducting the CRT after the experiment can have a negative effect on performance on the CRT[15].

3.5. Exposure to the CRT over the years (visibility)[16]

Toplak et al. (2014) argue that the test in its original form is becoming increasingly popular and is perhaps losing its efficacy. This argument has validity if the student pool remains the same, or same subjects take the test on more than one occasion over their University life. The critique is of concern given the increased implementation of the CRT and if we are to believe in its predictive power. This issue is also related with the fact that some studies are conducted on-line. Answers to the CRT are easily available online and this sheds doubt on its efficacy using online studies. We investigate these issues below.

Table 4 presents the number of studies included by year in our meta-analysis. In our regressions we used the variable *visibility* to describe the effect of exposure to the CRT over the years. The variable was generated by assigning the value 0 for studies conducted in 2005, 1 for 2006 and so on.

Table 4: Number of studies included according to the year they were conducted

Year of study	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Number of studies	1	6	3	4	15	16	15	27	15	16

Note: The sample does not include any CRT study from 2015.

[15] Poor performance may also be due to lack of effort, or leaving problems blank. We lack data on whether an incorrect answer is wrong or empty. We would like to thank Shane Frederick for pointing this out.

[16] This issue has been recently addressed by Stagnaro, Pennycook and Rand (2018, JDM) who have shown that CRT responses are stable across time.

In Table 3 (row 5) the variable *visibility* shows that the number of years of exposure has a positive impact on obtaining all three correct answers (*p<0.05*). *Visibility* negatively affects subjects answering only one question correctly *(p<0.01)*, the coefficients on two and three correct answers turn positive but non-significant. No effect is found for *None (p>0.1)*, i.e. exposure to the test is not decreasing the number of participants giving zero correct answers. In addition, we find that subjects are more likely to answer the *Machines* question correctly *(p<0.01)*. Overall, some support (row 5, Table 3) is lent to the argument that years of exposure positively affect test outcomes. This effect, however, does not seem to be too large or persistent. The robustness checks provide similar weak findings for the exposure conjecture for the females only subsample (Table ER1), for non-students (Table ER3), studies without economic experiments (Table ER4) and for studies excluding CRTs with random order (Table ER6). However, results show higher statistical power for the subsample of studies without monetary incentives (Table ER2).

The earlier results could be confounded by the presence of AMT studies in our sample. AMT studies have the potential problem of immediate internet access enabling easily access to the answers for the standard questions. Our results change when AMT studies are excluded from the sample (row 5, Table ER5). We find that the previously significant effects on *visibility* are substantially weaker. All in all, we cannot observe a clear link between length of exposure and obtaining correct answers in the CRT. This gives us the result below,

Result 5: Excluding the AMT studies we find little support for the fact that performance on the test improves as its exposure increases.

Besides our main hypotheses we also check for how using the standard test sequence, or hand-run vs computerized implementation impacts CRT performance. Below we present these results.

3.6. The sequence of questions

The most commonly used sequence for the CRT is the one originally proposed by Frederick (2005), i.e. *B&B*, *Machines*, and *Lillypad.* A large proportion of our sample, *83.78%,* corresponds to this. It is thus reasonable to see whether the standard implementation of the sequence affects final outcomes. Looking at all the studies (standardized and random sequences) we see that subjects score better on the CRT when the questions are presented in the standard order (Table 3). However, excluding studies with randomized and other forms of implementation we find that the effect of standardized implementation on CRT responses is marginal. We thus find no clear evidence on the effect of implementing the standard sequence upon outcomes.

Looking at both the standard and randomized studies we find that the coefficient of *standard* sequence (row 6 - Table 3) is significant for the *B&B (p<0.05)*, *Machines (p<0.01)* and *Lillypad (p<0.01)* questions. Further, the likelihood of *None* is much higher when the questions are not asked in the standard order *(p<0.01)*. Likewise, subjects are more likely to answer two *(p<0.01)* or three *(p<0.01)* questions for the standard implementation. This result is, however, not robust.

Note that the randomized sequences can also include questions asked in a standard way with probability 1 in 6. [17] Controlling for 'other sequences' and excluding studies where the order of the questions was randomized (11.64% of all of our observations) we find that the effect is marginal. In Table ER5 we replicated the main regressions excluding the studies with random sequences. The effect of standardized sequence on correct CRT responses is now marginal (Row 5, Table ER6). We cannot thus conclude that the standardized sequence would bias responses in the CRT.

3.7 Hand run vs. computerized?

Next we explore whether different administration modes effect performance on the test. In this case one would not expect that either method of implementation expects outcomes as it is a problem solving task not involving specific decisions (as occurs in most experiments). However, it is still interesting to study whether different forms of implementation affect final outcomes.

We find (Table 3, row 7) that the dummy variable for *computerized* is only weakly significant. We do observe that subjects using computers are less likely *(p<0.05)* to fail all three questions and more likely to have two correct answers *(p<0.05)*. Further, computerized implementation favors performance in the *Machines (p<0.1)* and *Lillypad* questions *(p<0.05)*, however, we do not observe significant effects on the *B&B* question *(p>0.1)*. We find this puzzling since one would expect that using paper and pencil would be more conducive to obtaining correct answers.

Our robustness check show similar but slightly more significant results for the subsample of females (Table ER1) and the studies without monetary incentives (Table ER2) and the subsample of studies

[17] If we consider that 1/6 of the randomized sample use standard sequence (roughly 2% of the sample) then we have that 85.7% of the sample uses the standard sequence and 14.3% non-standard (including 5/6 of the random).

excluding CRTs with random order (Table ER4), however, the results show less statistical power for the subsample of non-students (Table ER3). Finally, the subsample of studies without economic experiments (Table ER4) and the subsample excluding Amazon Mechanical Turk studies (ATM, Table ER5) produce similar results. Note, however, we do not have information on whether participants could work out solutions on paper while responding to the computerized questions. Summarizing, we find that running the CRT on computers as compared to paper and pencil results in weakly significant positive effects on test scores.

4. Discussion

The CRT has become increasingly popular in predicting behavior in Economics and Psychology experiments. However, there is no consensus on how the vastly different implementation procedures used, i.e. being incentivized or not, administered by paper-and-pencil/computers/AMT, before/in-the-middle/after-an-experiment, etc. impact performance on the CRT. We only know from Frederick (2005) that the test has a strong (male) gender bias. The purpose of this study is to provide the first extensive look at how different implementation procedures for CRT may impact performance on the test. In the end if the CRT is useful for its predictive power then knowing whether any small variation in implementation procedures can affect outcomes is important.

In this paper we conduct a meta-survey of the methods employed in 118 studies (N = 44,558) that use CRT. Our main result reaffirms and provides additional findings regarding the gender bias result first reported in Frederick (2005). We find that males perform notably better in this test. This observation is important if, say, one is interested in constructing samples based on cognitive ability. This could lead to strong (gender) sample imbalance. For instance, if one uses three correct answers as a selection criteria then the sample is disproportionately biased towards males. Our second interesting finding is that we find no statistical evidence to support the argument that *monetary incentives* may play an important role in improving CRT performance. Albeit limited (as we lack data on the amount, or how, subjects were paid), this result is important as it tells us that incentives may not be strongly relevant for the implementation of the CRT. Regarding comparing student vs non-student populations we find that *students* are more likely to answer all three questions correctly compared to non-students, and less likely to have zero correct answers. Again this tells us that the predictive power of the CRT may be affected by population differences.

We also find that conducting the CRT *after the experiments* negatively effects test outcomes. Conducting the test later decreases the probability of obtaining correct answers; meanwhile, the probability of obtaining *None* is increased. This result is interesting as it points towards the fact that increased cognitive load could be an important determinant of performance in the CRT. Another interpretation of this result could be that it provides indirect support to the argument that glucose is important in cognitive tasks and cognition declines with time and effort. This is important as after removing studies where the researchers did not run experiments from the data we find even more significant results. We test for the year effect (*visibility*) and find no clear evidence that exposure positively affects tests results.

Comparing test scores for hand-run vs. *computerized* tests we found a weakly positively significant effect of computerized implementation of the test. It is important to point out that we do not collect individual CRT scores but session information about CRT score distribution and do not control for individual characteristics such as cognitive ability, for example measured by IQ. This makes the analysis of individual characteristics challenging. Finally, we should add that, as is common with studies of this nature, a comprehensive list of data was not available. We lacked information about particular details (such as length of experiment, size of incentives, etc.) of each experiment in our meta-study. Knowing these details would have aided the interpretation of our results.

References

Aiken, L. (1986-1987). Sex differences in mathematical ability: A review of the literature. *Educational Research Quarterly, 10*, 25-35.

Alós-Ferrer, C., & Hügelschäfer, S. (2016). Faith in intuition and cognitive reflection. *Journal of Behavioral and Experimental Economics, 64*, 61–70.

Andersson, O., Holm, H. J., Tyran, J. R., & Wengström, E. (2016). Risk aversion relates to cognitive ability: Preferences or noise? *Journal of the European Economic Association*, *14*(5), 1129-1154.

Awasthi, V., & Pratt, J. (1990). The effects of monetary incentives on effort and decision performance: The role of cognitive characteristics. *The Accounting Review, 65(4)*, 797-811.

Benbow, C.P., & Stanley, J.C. (1980). Sex differences in mathematical ability: Fact or artifact? *Science, 210(4475)*, 1262–264.

Benbow, C. P., Lubinski, D., Shea, D. L., & Eftekhari-Sanjani, H. (2000). Sex differences in mathematical reasoning ability: Their status 20 years later. *Psychological Science, 11*, 474-480.

Ben-Ner, A., Kong, F., & Putterman, L. (2004). Share and share alike? Gender-pairing, personality, and cognitive ability as determinants of giving. *Journal of Economic Psychology, 25(5)*, 581-589.

Besedes, T., Deck, C., Sarangi, S., & Shor, M. (2012). Decision-making strategies and performance among seniors. *Journal of Economic Behavior and Organization, 81(2*), 524-533.

Bonner, S.E., & Sprinkle, G.B. (2002). The effects of monetary incentives on effort and task performance: theories, evidence, and a framework for research. *Accounting, Organizations and Society, 27*, 303–345.

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.

Bosch-Rosa, C., Meissner, T., & Bosch-Domenech, A. (2018). Cognitive bubbles. *Experimental Economics*, *21*(1), 132-153.

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

Brown, A. L., Viriyavipart, A. & Wang, X. (2018), Search deterrence in experimental consumer goods markets, *European Economic Review*, 104, 167-84.

Buhrmester, M., Kwang, T., & Gosling, S.D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6(1)*, 3-5.

Camerer, C. F., & Hogarth, R.M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty, 19(1)*, 7-42.

Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. *Judgment and Decision Making, 5(3)*, 182-191.

Carpenter, J., Graham, M., & Wolf, J. (2013). Cognitive ability and strategic sophistication. *Games and Economic Behavior, 80(1)*, 115–130.

Chen, C.C., Chiu, I.M., Smith, J., & Yamada, T. (2013). Too smart to be selfish? Measures of cognitive ability, social preferences, and consistency. *Journal of Economic Behavior & Organization, 90(0)*, 112–122.

Cheung, S. L., Hedegaard, M., & Palan, S. (2014). To see is to believe. Common expectations in experimental asset markets. *European Economic Review, 66*, 84–96.

Cole, M.S., Bedeian, A.G., & Field, H.S. (2006). The measurement equivalence of web-based and paper-and-pencil measures of transformational leadership: A multinational test. *Organizational Research Methods, 9(3)*, 339-368.

Corgnet, B., Espin, A.M., Hernan-Gonzalez, R., Kujal, P., & Rassenti, S. (2016). To trust, or not to trust: Cognitive reflection in trust games. *Journal of Behavioral & Experimental Economics, 64*, 20-27.

Corgnet, B., Espín, A.M., & Hernán-González, R. (2015). The cognitive basis of social behavior: cognitive reflection overrides antisocial but not always prosocial motives. *Frontiers in Behavioral Neuroscience, 9(287).*

Corgnet, B., Hernan-Gonzalez, R., Kujal, P., & Porter, D. (2014). The effect of earned versus house money on price bubble formation in experimental asset markets, *Review of Finance*, 19: 1455-1488.

Cueva-Herrero, C., Iturbe-Ormaetxe, I., Mata-Prez, E., Ponti, G., Sartarelli, M., Yu, H., & Zhukova, V. (2016). Cognitive (ir)reflection: New experimental evidence. *Journal of Behavioral and Experimental Economics*, *64*, 81-93.

Diamond, D.W., & Dybvig, P.H. (1983). Bank runs, deposit insurance, and liquidity. *Journal of Political Economy, 91(3)*, 401-419.

Duckworth, A.L., Quinn, P.D., Lynam, D.R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences, 108*, 7716–7720.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, *49*, 709–724.

Exadaktylos, F., Espín, A.M., & Branas-Garza, P. (2013). Experimental subjects are not different. *Scientific Reports*, *3(1213*), 1-6.

Falk, A., & Heckman, J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science, 326*, 535–538.

Falk, A., Meier, S., & Zehnder, C. (2013). Do lab experiments misrepresent social preferences? The case of self-selected student samples. *Journal of the European Economic Association, 11(4)*, 839–852.

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

Gailliot, M.T., & Baumeister, R.F. (2007). The physiology of willpower: Linking blood glucose to self-control. *Personality and Social Psychology Review, 11(4)*, 303-327.

George, C.E., Lankford, J.S. & Wilson, S.E. (1992). The effects of computerised versus paper-and-pencil administration on measures of negative affect. *Computers in Human Behavior, 8(2-3)*, 203-209.

Gideon, N., Nadler, A., Zava, D. & Camerer, C.F. (2017), Single-dose testosterone administration impairs cognitive reflection, *Psychological Science* 28, 1398-1407.

Gneezy, U., & Rustichini, A. (2000a). Pay enough or don't pay at all. *The Quarterly Journal of Economics, 115(3)*, 791–810.

Goodman, J. K., Cryder, C.E., & Cheema, A.A. (2013). Data collection in a flat world: strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*, 213-224.

Hanaki N., Jacquemet N., Luchini S. & Zylbersztejn A. (2016) Fluid intelligence and cognitive reflection in a strategic environment: evidence from dominance-solvable games, *Frontiers in Psychology*, 2016, 7:1188.

Holt, C.A, Porzio, M., & Song, M.Y. (2015). Price bubbles, gender, and expectations in experimental asset markets. *European Economic Review, 100, 72-94.*

Hoppe, E. I., & Kusterer, D.J. (2011). Behavioral biases and cognitive reflection. *Economics Letters, 110,* 97–100.

Jenkins, G. D., Mitra, A., Gupta, N., & Shaw, J.D. (1998). Are financial incentives related to performance? A meta-analytic review of empirical research. *Journal of Applied Psychology, 83(5)*, 777-787.

Kahneman, D., & Frederick, S. (2002) Representativeness revisited: At- tribute substitution in intuitive judgment. In T. Gilovich, D. Griffin and D. Kahneman (Eds.), Heuristics and biases: The psychology of intuitive judgment, 49-81, New York: Cambridge University Press.

King, W.C., & Miles, E.W. (1995). A quasi-experimental assessment of the effect of computerizing noncognitive paper-and-pencil measurements: A test of measurement equivalence. *Journal of Applied Psychology, 80*, 643-651.

Kiss, H.J., Rodriguez-Lara, I., & Rosa-García, A. (2016). Think twice before running! Bank runs and cognitive abilities. *Journal of Behavioral and Experimental Economics, 64*, 12-19.

Königsheim, C., Lukas, M., & Nöth, M. (2018) Individual preferences and the exponential growth bias, Journal of Economic Behavior & Organization, 145: 352-69.

Königsheim, C., Lukas, M., & Nöth, M. (2019), Salience theory: Calibration and heterogeneity in probability distortion, *Journal of Economic Behavior & Organization*, 2019, 157, 477-95.

Levitt, S.D., & List, J.A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives, 21(2)*, 153–174.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making, 25(4)*, 361–381.

Lukas, M., and Markus Nöth (2019), Interest Rate Changes and Borrower Search Behavior, *Journal of Economic Behavior & Organization*, 163, 172-189.

Mau, W.C., & Lynn, R. (2010). Gender differences in homework and test scores in mathematics, reading and science at tenth and twelfth grade. *Psychology, Evolution & Gender*, *2(2)*, 119-125.

Moritz, B., Hill, A.V., & Donohue, K. (2013). Individual differences in the newsvendor problem: Behavior and cognitive reflection. *Journal of Operations Management, 31(1-2)*, 72-85.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review*, *85(5)*, 1313-26.

Neyse L., Bosworth S., Ring P. and Schmidt, U. (2016). Overconfidence, Incentives and Digit Ratio, *Scientific Reports*, 6: 23294.

Noussair, C.N., Trautmann, S.T., & van de Kuilen G. (2014). Higher order risk attitudes, demographics, and financial decisions. *The Review of Economic Studies*, *81 (1)*, 325-355.

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

Paolacci, G., Chandler, J., & Ipeirotis, P.G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5(5)*, 411–419.

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

Peterson, R.A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research, 28(3)*, 450-461.

Peysakhovich, A., & Rand, D.G. (2015). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science,* doi:10.1287/mnsc.2015.2168

Pikulina, Elena, Renneboog, Luc and Tobler, Philippe N. (2017), Overconfidence and investment: An experimental approach, *Journal of Corporate Finance*, 104, 167-84.

Ponti, G., & Rodriguez-Lara, I. (2015). Social preferences and cognitive reflection: Evidence from dictator game experiment. *Frontiers in Behavioral Neuroscience,* doi: 10.3389/fnbeh.2015.00146

Ruffle, Bradley J., and Wilson, Anne E. (2018), The truth about tattoos. *Economics Letters*, 172, 143-47.

Riedel, J.A., Nebeker, D.M., & Cooper, B.L. (1988). The influence of monetary incentive on goal choice, goal commitment, and task performance. *Organizational Behavior and Human Decision Processes, 42*, 155-180.

Scott, W. E., Farh, J.L., & Podsakoff, P.M. (1988). The effects of ‘intrinsic’ and ‘extrinsic’ reinforcement contingencies on task behavior. *Organizational Behavior and Human Decision Processes*, *41*, 405-425.

Sheremeta, R. M., (2018). Impulsive behavior in competition: Testing theories of overbidding in rent-seeking contests, *ESI Working paper* 2018.

Sloman, S.A. (1996) The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.

Stagnaro MN, Pennycook G, Rand DG (2018) Performance on the Cognitive Reflection Test is stable across time. *Judgement and Decision Making*, 13, 3, 260-267.

Stanovich, K.E., & West, R.F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*, 645-665.

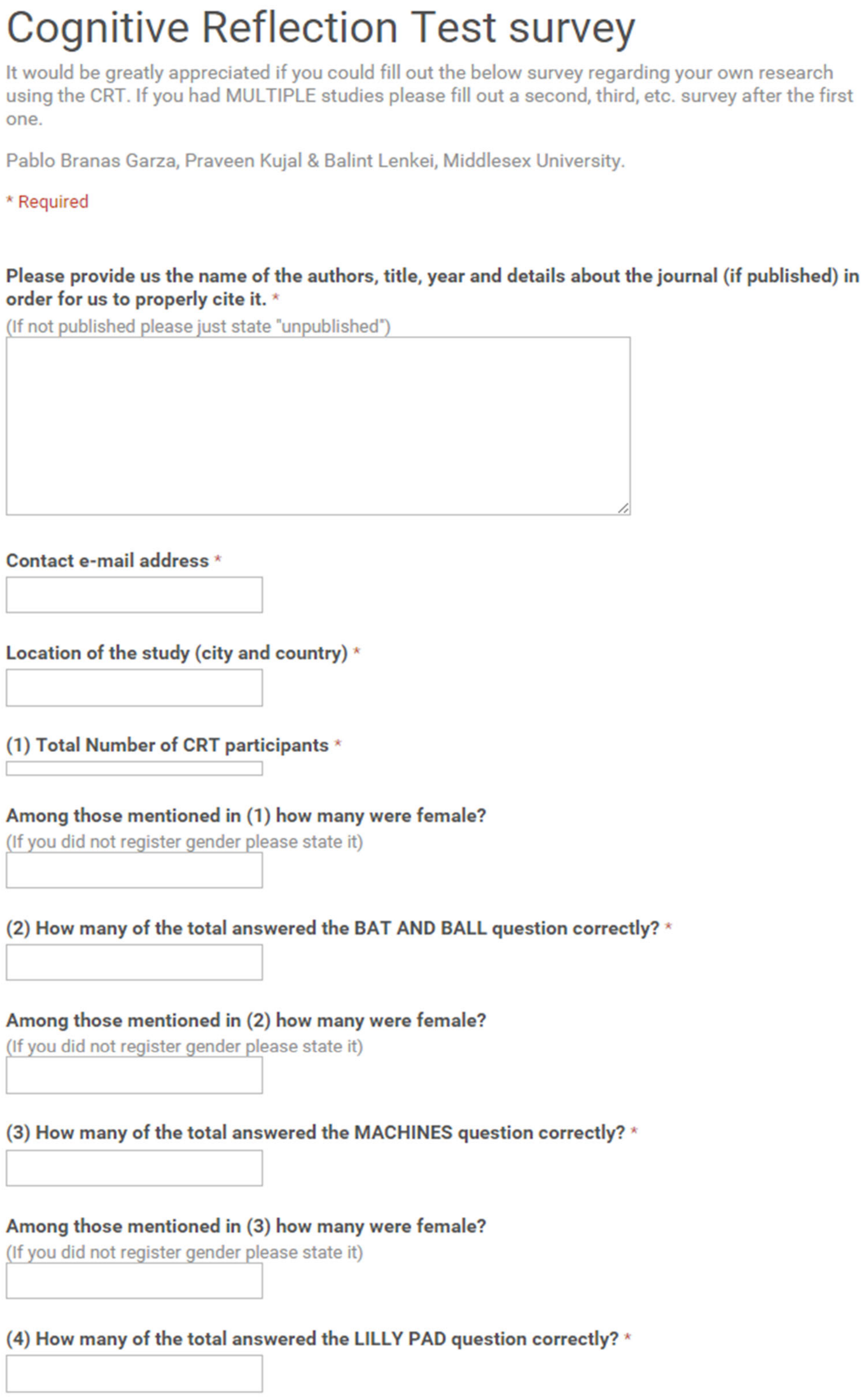
Toplak, M.E., West, R.F., & Stanovich, K.E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics and biases tasks. *Memory & Cognition*, *39*, 1275-1289.

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.

Weiss, V. (1986). From memory span to the quantum mechanics of intelligence. *Personality and Individual Differences*, *7*, 737-749.

Appendix A

Figure A1. Screenshot of the Cognitive Reflection Test survey



Cognitive Reflection Test survey

It would be greatly appreciated if you could fill out the below survey regarding your own research using the CRT. If you had MULTIPLE studies please fill out a second, third, etc. survey after the first one.

Pablo Branas Garza, Praveen Kujal & Balint Lenkei, Middlesex University.

* Required

Please provide us the name of the authors, title, year and details about the journal (if published) in order for us to properly cite it. *

(If not published please just state "unpublished")

Contact e-mail address *

Location of the study (city and country) *

(1) Total Number of CRT participants *

Among those mentioned in (1) how many were female?

(If you did not register gender please state it)

(2) How many of the total answered the BAT AND BALL question correctly? *

Among those mentioned in (2) how many were female?

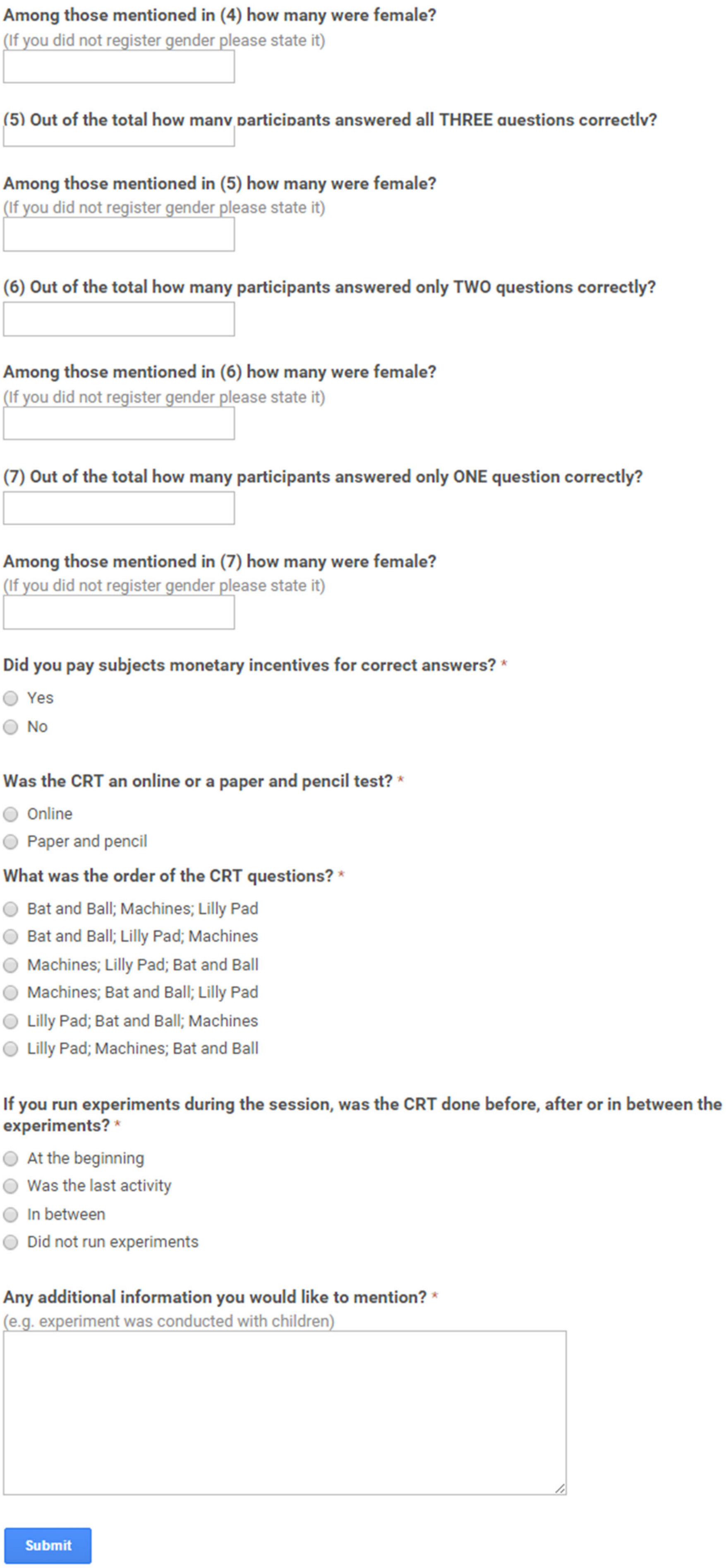
(If you did not register gender please state it)

(3) How many of the total answered the MACHINES question correctly? *

Among those mentioned in (3) how many were female?

(If you did not register gender please state it)

(4) How many of the total answered the LILLY PAD question correctly? *



Among those mentioned in (4) how many were female?

(If you did not register gender please state it)

(5) Out of the total how many participants answered all THREE questions correctly?

Among those mentioned in (5) how many were female?

(If you did not register gender please state it)

(6) Out of the total how many participants answered only TWO questions correctly?

Among those mentioned in (6) how many were female?

(If you did not register gender please state it)

(7) Out of the total how many participants answered only ONE question correctly?

Among those mentioned in (7) how many were female?

(If you did not register gender please state it)

Did you pay subjects monetary incentives for correct answers? *

○ Yes

○ No

Was the CRT an online or a paper and pencil test? *

○ Online

○ Paper and pencil

What was the order of the CRT questions? *

○ Bat and Ball; Machines; Lilly Pad

○ Bat and Ball; Lilly Pad; Machines

○ Machines; Lilly Pad; Bat and Ball

○ Machines; Bat and Ball; Lilly Pad

○ Lilly Pad; Bat and Ball; Machines

○ Lilly Pad; Machines; Bat and Ball

If you run experiments during the session, was the CRT done before, after or in between the experiments? *

○ At the beginning

○ Was the last activity

○ In between

○ Did not run experiments

Any additional information you would like to mention? *

(e.g. experiment was conducted with children)

Table ER1. Robustness check: Females only

	(1) *B&B*	(2) *Machines*	(3) *Lillypad*	(4) *None*	(5) *1*	(6) *2*	(7) *3*
(1) monetary incentives	0.012	-0.064	0.101*	0.000	-0.016	-0.037	0.053
	(0.057)	(0.064)	(0.053)	(0.066)	(0.021)	(0.030)	(0.034)
(2) student	0.127***	-0.027	0.039	-0.080**	0.031***	0.029	0.020
	(0.036)	(0.026)	(0.041)	(0.039)	(0.009)	(0.018)	(0.022)
(3a) in-between experiments	-0.046	0.009	-0.071	0.047	0.000	-0.015	-0.032
	(0.047)	(0.039)	(0.053)	(0.048)	(0.014)	(0.020)	(0.030)
(3b) after the experiment	-0.045	-0.004	-0.092**	0.064	-0.010	-0.029	-0.025
	(0.043)	(0.035)	(0.042)	(0.043)	(0.010)	(0.018)	(0.026)
(4) visibility	0.009*	0.017***	0.008	-0.007	-0.007***	0.004	0.011**
	(0.005)	(0.005)	(0.005)	(0.005)	(0.002)	(0.002)	(0.004)
(5) standard sequence	0.093**	0.106***	0.151***	-0.149***	0.017	0.059***	0.072**
	(0.041)	(0.033)	(0.044)	(0.044)	(0.013)	(0.018)	(0.028)
(6) computerized	0.069*	0.100**	0.112**	-0.130***	0.032**	0.058***	0.040
	(0.038)	(0.048)	(0.045)	(0.049)	(0.015)	(0.022)	(0.025)
constant	0.053	0.072	0.039	0.759***	0.228***	0.086**	-0.073*
	(0.069)	(0.073)	(0.068)	(0.080)	(0.027)	(0.036)	(0.044)
N	19995	19995	19995	20945	20945	20945	20945
R^2	0.026	0.020	0.032	0.031	0.005	0.009	0.013

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, *p<0.1. The regressions also controls for the country of the study by using two dummy variables: europe and anglo-saxon.

Table ER2. Robustness check: excluding studies where the experimenters used monetary incentives to reward correct answers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	B&B	*Machines*	*Lillypad*	*None*	*1*	*2*	*3*
(1) female	-0.107***	-0.176***	-0.197***	0.181***	0.004	-0.066***	-0.118***
	(0.011)	(0.010)	(0.011)	(0.010)	(0.007)	(0.008)	(0.008)
(2) student	0.108***	-0.037	0.046	-0.066*	0.017**	0.027*	0.022
	(0.036)	(0.024)	(0.040)	(0.035)	(0.007)	(0.014)	(0.023)
(3a) in-between experiments	-0.070	-0.055*	-0.115**	0.083*	0.012	-0.026	-0.069**
	(0.050)	(0.030)	(0.056)	(0.045)	(0.010)	(0.016)	(0.032)
(3b) after the experiment	-0.065	-0.047	-0.123***	0.088**	-0.004	-0.033**	-0.051*
	(0.040)	(0.033)	(0.043)	(0.038)	(0.008)	(0.014)	(0.027)
(4) visibility	0.016***	0.024***	0.011**	-0.011**	-0.009***	0.004	0.016***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.002)	(0.002)	(0.004)
(5) standard sequence	0.119***	0.122***	0.162***	-0.153***	0.006	0.053***	0.094***
	(0.039)	(0.032)	(0.041)	(0.040)	(0.011)	(0.015)	(0.029)
(6) computerized	0.063*	0.112**	0.154***	-0.132***	0.016	0.066***	0.051
	(0.036)	(0.047)	(0.044)	(0.040)	(0.013)	(0.018)	(0.032)
constant	0.271***	0.390***	0.365***	0.428***	0.246***	0.177***	0.149**
	(0.088)	(0.073)	(0.084)	(0.086)	(0.026)	(0.039)	(0.062)
N	31766	31766	31766	33338	33338	33338	33338
R^2	0.051	0.063	0.077	0.072	0.005	0.016	0.046

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, *p<0.1. The regressions also controls for the country of the study by using two dummy variables: europe and anglo-saxon.

Table ER3. Robustness check: Regressions with non-student samples only

	(1) *B&B*	(2) *Machines*	(3) *Lillypad*	(4) *None*	(5) *1*	(6) *2*	(7) *3*
(1) female	-0.086***	-0.154***	-0.178***	0.165***	-0.007	-0.061***	-0.097***
	(0.010)	(0.013)	(0.014)	(0.012)	(0.008)	(0.010)	(0.008)
(2) monetary incentives	0.073	0.066	0.189**	-0.108	-0.023**	0.058***	0.072
	(0.090)	(0.097)	(0.070)	(0.088)	(0.008)	(0.020)	(0.073)
(3a) in-between experiments	-0.051	-0.044	-0.093*	0.077*	0.017**	-0.026*	-0.069*
	(0.046)	(0.031)	(0.048)	(0.039)	(0.008)	(0.014)	(0.035)
(3b) after the experiment	-0.032	-0.004	-0.049	0.045	-0.003	-0.011	-0.031
	(0.023)	(0.048)	(0.042)	(0.042)	(0.007)	(0.018)	(0.028)
(4) visibility	0.040***	0.028	0.018	-0.022	-0.010***	0.003	0.029***
	(0.006)	(0.017)	(0.015)	(0.014)	(0.002)	(0.006)	(0.009)
(5) standard sequence	0.189***	0.156***	0.199***	-0.203***	-0.004	0.061***	0.146***
	(0.038)	(0.038)	(0.044)	(0.040)	(0.012)	(0.016)	(0.031)
(6) computerized	-0.030	0.012	0.053	-0.057	0.041***	0.035	-0.020
	(0.065)	(0.066)	(0.057)	(0.063)	(0.005)	(0.022)	(0.051)
constant	-0.193**	0.088	0.006	0.817***	0.278***	0.085**	-0.180**
	(0.073)	(0.111)	(0.095)	(0.097)	(0.016)	(0.037)	(0.069)
N	21983	21983	21983	23199	23199	23199	23199
R^2	0.041	0.044	0.078	0.071	0.007	0.017	0.042

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, *p<0.1. The regressions also controls for the country of the study by using two dummy variables: europe and anglo-saxon.

Table ER4. Robustness check: excluding the studies where the researchers did not run experiments

	(1) *B&B*	(2) *Machines*	(3) *Lillypad*	(4) *None*	(5) *1*	(6) *2*	(7) *3*
(1) female	-0.107***	-0.167***	-0.186***	0.170***	0.009	-0.063***	-0.116***
	(0.012)	(0.008)	(0.010)	(0.010)	(0.008)	(0.007)	(0.009)
(2) monetary incentives	0.046	0.091*	0.110*	-0.081	-0.007	0.019	0.069*
	(0.045)	(0.051)	(0.059)	(0.052)	(0.017)	(0.021)	(0.039)
(3) student	0.108***	-0.026	0.051	-0.070	0.014	0.030*	0.026
	(0.037)	(0.029)	(0.049)	(0.042)	(0.009)	(0.017)	(0.024)
(4a) in-between experiments	-0.130**	-0.039	-0.140**	0.101**	0.015	-0.022	-0.093**
	(0.055)	(0.048)	(0.058)	(0.050)	(0.015)	(0.020)	(0.036)
(4b) after the experiment	-0.109***	-0.037	-0.135***	0.095**	0.005	-0.029	-0.071**
	(0.041)	(0.048)	(0.047)	(0.041)	(0.012)	(0.019)	(0.028)
(5) visibility	0.002	0.013*	0.005	-0.004	-0.004	0.004	0.004
	(0.007)	(0.007)	(0.009)	(0.008)	(0.003)	(0.003)	(0.006)
(6) standard sequence	0.120***	0.115***	0.175***	-0.164***	0.018	0.057***	0.089***
	(0.038)	(0.031)	(0.041)	(0.038)	(0.012)	(0.015)	(0.029)
(7) computerized	0.084**	0.136***	0.145***	-0.130***	0.000	0.055***	0.076***
	(0.035)	(0.038)	(0.048)	(0.041)	(0.010)	(0.021)	(0.024)
constant	0.392***	0.441***	0.396***	0.399***	0.195***	0.170***	0.237***
	(0.117)	(0.100)	(0.119)	(0.114)	(0.033)	(0.047)	(0.082)
N	28268	28268	28268	28624	28624	28624	28624
R^2	0.056	0.068	0.086	0.086	0.002	0.019	0.048

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, *p<0.1. The regressions also controls for the country of the study by using two dummy variables: europe and anglo-saxon.

Table ER5. Robustness check: excluding those studies where the experimenters used Amazon Mechanical Turk for the tests

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Bat and Ball	*Machines*	*Lillypad*	*None*	*1*	*2*	*3*
(1) female	-0.115***	-0.181***	-0.202***	0.180***	0.014**	-0.070***	-0.124***
	(0.012)	(0.011)	(0.011)	(0.011)	(0.007)	(0.008)	(0.009)
(2) monetary incentives	-0.022	0.006	0.045	-0.010	-0.002	0.002	0.010
	(0.043)	(0.042)	(0.047)	(0.043)	(0.014)	(0.017)	(0.036)
(3) student	0.171***	0.033	0.095**	-0.113***	0.001	0.031**	0.081***
	(0.041)	(0.032)	(0.046)	(0.042)	(0.009)	(0.015)	(0.029)
(4a) in-between experiments	-0.033	0.019	-0.088*	0.054	-0.010	-0.020	-0.023
	(0.047)	(0.041)	(0.049)	(0.042)	(0.015)	(0.016)	(0.035)
(4b) after the experiment	-0.030	-0.001	-0.093*	0.055	-0.008	-0.025	-0.022
	(0.045)	(0.035)	(0.047)	(0.043)	(0.009)	(0.015)	(0.032)
(5) visibility	0.003	0.010*	0.001	-0.002	-0.005**	0.002	0.005
	(0.006)	(0.006)	(0.006)	(0.005)	(0.002)	(0.002)	(0.005)
(6) standard sequence	0.059	0.059	0.118**	-0.121***	0.042**	0.042**	0.038
	(0.041)	(0.036)	(0.046)	(0.045)	(0.016)	(0.019)	(0.029)
(7) computerized	0.032	0.084*	0.106**	-0.095**	0.014	0.049**	0.032
	(0.040)	(0.049)	(0.052)	(0.046)	(0.011)	(0.020)	(0.034)
constant	0.248***	0.339***	0.333***	0.499***	0.200***	0.167***	0.134**
	(0.077)	(0.075)	(0.079)	(0.081)	(0.025)	(0.035)	(0.054)
N	31200	31200	31200	31870	31870	31870	31870
R^2	0.049	0.057	0.068	0.064	0.003	0.013	0.043

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, *p<0.1. The regressions also controls for the country of the study by using two dummy variables: europe and anglo-saxon.

Table ER6. Robustness check: Excluding studies where the sequence of questions was randomized

	(1) *B&B*	(2) *Machines*	(3) *Lillypad*	(4) *None*	(5) *1*	(6) *2*	(7) *3*
(1) female	-0.117***	-0.176***	-0.196***	0.176***	0.012**	-0.065***	-0.124***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.034)	(0.000)	(0.000)
(2) computerized	0.021	0.077	0.098*	-0.087*	0.013	0.048**	0.026
	(0.601)	(0.118)	(0.062)	(0.060)	(0.274)	(0.017)	(0.443)
(3) student	0.111***	-0.017	0.046	-0.063*	0.011	0.022*	0.030
	(0.003)	(0.520)	(0.272)	(0.079)	(0.183)	(0.083)	(0.251)
(4a) in-between experiments	-0.055	0.000	-0.097*	0.064	-0.003	-0.019	-0.043
	(0.266)	(0.995)	(0.073)	(0.147)	(0.845)	(0.231)	(0.242)
(4b) after the experiment	0.007	0.027	-0.050	0.017	-0.010	-0.013	0.006
	(0.859)	(0.371)	(0.193)	(0.606)	(0.236)	(0.305)	(0.826)
(5) standard sequence	-0.031	-0.087*	-0.044	0.024	0.047*	0.001	-0.072**
	(0.524)	(0.068)	(0.288)	(0.664)	(0.092)	(0.956)	(0.016)
(6) monetary incentives	-0.005	0.019	0.060	-0.025	-0.003	0.005	0.023
	(0.918)	(0.690)	(0.224)	(0.584)	(0.867)	(0.772)	(0.570)
(7) visibility	0.007	0.014**	0.004	-0.004	-0.007***	0.002	0.009*
	(0.214)	(0.023)	(0.532)	(0.424)	(0.004)	(0.281)	(0.096)
constant	0.317***	0.460***	0.475***	0.373***	0.204***	0.201***	0.222***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
N	32846	32846	32846	34418	34418	34418	34418
R^2	0.037	0.049	0.053	0.048	0.003	0.01	0.036

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, *p<0.1. The regressions also controls for the country of the study by using two dummy variables: europe and anglo-saxon.

Appendix B

List of articles included in the meta-study[18,19]

#001 Agranov, M., Caplin, A., Tergiman, C. (2015) Naive Play and the Process of Choice in the Guessing Games. Forthcoming in Journal of the Economics Science Association.

#002 Akiyama, E., Hanaki, N., Ishikawa, R. (2013) It is not just confusion! Strategic uncertainty in an experimental asset market, Aix-Marseille School of Economics Working Paper 2013 No. 40.

#003 Akiyama, E., Hanaki, N., Ishikawa, R. (2014) How do experienced traders respond to inflows of inexperienced traders? An experimental analysis. Journal of Economic Dynamics and Control, 45(C): 1-18.

#004 Alós-Ferrer, C., Hügelschäfer, S. (2014) Faith in Intuition and Cognitive Reflection. University of Cologne Working Paper. Study 3

#005 Alter, A. L., Oppenheimer, D. M., Epley, N., Eyre, R. N. (2007) Overcoming intuition: Metacognitive difficulty activates analytic reasoning. Journal of Experimental Psychology: General, 136: 569-576.

#006 Andersson, O., Tyran, J.R., Wengström, E., Holm, H.J. (2013) Risk Aversion Relates to Cognitive Ability: Fact or Fiction?. IFN Working Paper No. 964.

#007 Baghestanian, S., Frey, S. (2013) Go Figure. Analytic and Strategic Skills are Separable. Indiana University, Working paper.

#008 Balafoutas, L., Kerschbamer, R., Oexl, R. (2014) Distributional preferences and ego depletion. Working Paper.

#009 Barham, B., Chavas, J.P., Fitz, D., Salas, V.R., Schechter, L. (2014) The Roles of Risk and Ambiguity in Technology Adoption. Journal of Economic Behavior and Organization, 97: 204-218.

#010 Barr, N., Pennycook, G., Stolz, J.A., Fugelsang, J.A. (2015) Reasoned connections: A dual-process perspective on creative thought. Thinking & Reasoning, 21(1): 61-75.

#011 Barr, N., Pennycook, G., Stolz, J.A., Fugelsang, J.A. (2015) The brain in your pocket: Evidence that Smartphones are used to supplant thinking. Computers in Human Behavior, 48: 473-480.

#012 Bartling, B., Engl, F., Weber, R.A. (2014) Game Form Misconceptions Do Not Explain the Endowment Effect, CESifo Working Paper No. 5094.

[18] The papers listed in Appendix B appear as when they were submitted to be included in our meta-study.

[19] In 2018 we run a second call to disentangle the effect of incentives (and included new items in the survey). We received very few studies. These papers are listed as #106 to #114 but not included in the statistical analysis.

#013 Belot, M., Bhaskar, V., Van De Ven, J. (2012) Can Observers Predict Trustworthiness? Review of Economics and Statistics, 94:1: 246-259.

#014 Bergman, O., Ellingsen, T., Johannesson, M., Svensson, C. (2010) Anchoring and cognitive ability. Economics Letters, 107:66-68.

#015 Besedes, T., Deck, C., Quintanar, S.M., Sarangi, S., Shor, M. (2014) Effort and Performance: What Distinguishes Interacting and Non-interacting Groups from Individuals?. Southern Economic Journal, 81(2): 294-322.

#016 Besedes, T., Deck, C., Sarangi, S., Shor, M. (2012) Decision-making Strategies and Performance among Seniors. Journal of Economic Behavior and Organization, 81(2): 524-533.

#017 Bigoni, M., Dragone, D. (2012) Effective and efficient experimental instructions. Economics Letters, 117(2): 460-463.

#018 Borghans, L., Golsteyn, B.H.H. (2014) Default Options and Training Participation. Empirical Economics, 46(4): 1417-1428.

#019 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.

#020 Bosch-Rosa, C., Meissner, T., Bosch-Domènech, A. (2015) Cognitive bubbles. Universitat Pompeu Fabra, Department of Economics and Business Working Paper 1464.

#021 Brañas-Garza, P., García-Muñoz, T., Hernán-Gonzalez, R. (2012) Cognitive Effort In The Beauty Contest Game. Journal Of Economic Behavior And Organization, 83(2): 254-260.

#022 Browne, M., Pennycook, G., Goodwin, B., McHenry, M. (2014) Reflective minds and open hearts: Cognitive style and personality predict religiosity and spiritual thinking in a community sample. European Journal of Social Psychology, 44(7): 736–742.

#023 Camilleri, A. R., Larrick, R. P. (2014) Metric and scale design as choice architecture tools. Journal of Public Policy & Marketing, 33(1): 108-125.

#024 Campitelli, G., Gerrans, P. (2014) What does the cognitive reflection test measure? A mathematical modelling approach. Memory and Cognition, 42(3): 434-447.

#025 Campitelli, G., Labollita, M. (2010) Correlations of cognitive reflection with judgments and choices. Judgment and Decision Making, 5(3): 182-191.

#026 Carpenter, J., Graham, M., Wolf, J. (2013) Cognitive Ability and Strategic Sophistication. Games and Economic Behavior, 80(1): 115-130.

#027 Caudek, C. (2014) Individual differences in cognitive control on self-referenced and other-referenced memory. Consciousness and Cognition, 30: 169-183.

#028 Cheung, S. L., Hedegaard, M., Palan, S. (2014) To See is to Believe. Common Expectations in Experimental Asset Markets. European Economic Review, 66: 84–96.

#029 Cheyne, J.A., Pennycook, G. (2013) Sleep paralysis post-episode distress: Modeling potential effects of episode characteristics, general psychological distress, beliefs, and cognitive style. Clinical Psychological Science, 1: 135-148.

#030 Coates, E.L., Blaszczynski, A. (2014) Predictors of Return Rate Discrimination in Slot Machine Play. Journal of Gambling Studies, 30(3): 669-83.

#031 Cobo-Reyes, R., Jimenez, N. (2012) The dark side of friendship: 'envy'. Experimental Economics, 15(4): 547-570.

#032 Corgnet, B., Espin, A., Hernan-Gonzalez, R., Kujal, P., Rassenti, S. (2014) To trust, or not to trust: Cognitive reflection in trust games. Forthcoming in Journal of Behavioral & Experimental Economics.

#033 Corgnet, B., Hernán-Gonzalez, R., Kujal, P., Porter, D. (2014) The effect of earned vs. house money on price bubble formation in experimental asset markets. Review of Finance, rfu031.

#034 Costa, A., Foucart A., Hayakawa S., Aparici M., Apesteguia J., Heafner J. et al. (2014) Your morals depend on language. PLoS One, 9(4): 1-7.

#035 Drouvelis, M., Jamison, J.C. (2015) Selecting Public Goods Institutions: Who Likes to Punish and Reward?. Forthcoming in Southern Economic Journal.

#036 Duttle, K. (2015) Cognitive skills and confidence: Interrelations with overestimation, overplacement and overprecision. Working paper.

#037 Duttle, K., Inukai, K. (2015) Complexity Aversion: Influences of Cognitive Abilities, Culture and System of Thought. Economics Bulletin, 35(2): 846-855.

#038 Duttle, K., Shichijo, T. (2015) Default or Reactance? Identity Priming Effects on Overconfidence in Germany and Japan. Working paper.

#039 Fehr, D., Huck, S. (2013) Who knows it is a game? On strategic awareness and cognitive ability. WZB Discussion Paper SP II 2013-306. Berlin: WZB.

#040 Fosgaard, T. R., Hansen, L. G., & Piovesan, M. (2013). Separating Will from Grace: an experiment on conformity and awareness in cheating. Journal of Economic Behavior & Organization, 93, 279-284. 10.1016/j.jebo.2013.03.027

#041 Gómez-Chacón, I. M., García-Madruga, J.A., Vila, J.O., Elosúa, M.R., Rodríguez, R. (2014) The dual processes hypothesis in mathematics performance: Beliefs, Cognitive Reflection, Reasoning and Working Memory, Learning and Individual Differences, (29): 67–73.

#042 Goodman, J. K., Cryder, C.E., Cheema, A.A. (2013) Data Collection in a Flat World: Strengths and Weaknesses of Mechanical Turk Samples. Journal of Behavioral Decision Making, 26: 213-224.

#043 Gsottbauer, E., Bergh, J.C.J.M. van den (2014) Disaster perception and the likelihood of cooperation in mitigating climate change: An experimental analysis. Submitted to Climatic Change.

#044 Guillen, P., Rustamdjan, H. (2014) Monkey see, monkey do: truth-telling in matching mechanism and the manipulation of others. University of Sydney Working Papers.

#045 Guthrie, C., Rachlinski, J.J., Wistrich, A.J. (2008) Blinking On The Bench: How Judges Decide Cases. Cornell Law Review, 93(1): 1-44.

#046 Haita, C. (2013) Sunk-Cost Fallacy with Partial Reversibility: An Experimental Investigation. University of Hamburg, Working Paper 2013 No. 09.

#047 Haita, C. (2014) Sunk-Cost Fallacy and Cognitive Ability in Individual Decision-Making. Working Paper.

#048 Hanaki, N., Jacquemet, N., Luchini, S., Zylbersztejn, A. (2014) Cognitive Ability and the Effect of Strategic Uncertainty. AMSE Working Paper 2014-58.

#049 Haran, U., Ritov, I., Mellers, B. A. (2013) The role of actively open-minded thinking in information acquisition, accuracy, and calibration. Judgment and Decision Making, 8(3): 188-201.

#050 Hardisty, D.J., Weber, E.U. (2009) Discounting future green: Money vs the environment. Journal of Experimental Psychology: General, 138(3): 329-340.

#051 Herz, H., Taubinsky, D. (2014) What Makes a Price Fair? An Experimental Study of Market Experience and Endogenous Fairness Norms. University of Zurich, Department of Economics Working Paper No. 128.

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

#053 Hyejin, K., Salmon, T.C. (2012) The Incentive Effects of Inequality: An Experimental Investigation. Southern Economic Journal, 79(1): 46-70.

#054 Ibanez, M., Riener, G., Rai, A. (2013) Sorting through Affirmative Action: two field experiments in Colombia. Courant Research Centre: Poverty, Equity and Growth Working paper, University of Goettingen No. 150.

#055 Insler, M., Compton, J., Schmitt, P. (2013) Does Everyone Accept a Free Lunch? Decision Making Under (almost) Zero Cost Borrowing. Research in Experimental Economics, 16: 145 – 170.

#056 Kahan, D.M. (2013) Ideology, Motivated Reasoning, and Cognitive Reflection. Judgment and Decision Making, 8(4): 407-424.

#057 Kahan, D.M. (2015) Climate Science Communication and the Measurement Problem. Advances in Political Psychology, 36: 1–43.

#058 Kenju, K. (2014) Democracy and Resilient Pro-Social Behavioral Change: An Experimental Study. Working Paper.

#059 Kessler, J. B., Meier, S. (2014) Learning from (Failed) Replications: Cognitive Load Manipulations and Charitable Giving. Journal of Economic Behavior and Organization, 102: 10-13.

#060 Kinnunen, S. (2015) Sadism promotes altruism on the Internet. University of Helsinki Working Paper.

#061 Kinnunen, S.P., Lindeman, M., Verkasalo, M. (2014) Altruism on the Internet. University of Helsinki Working Paper.

#062 Kinnunen, S.P., Windmann, S. (2013) Dual-processing altruism. Frontiers in Psychology, 4: 1–8.

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

#064 Knobe, J., Samuels R. (2013) Thinking like a scientist: innateness as a case study. Cognition, 126(1): 72-86.

#065 Kocher, M.G., Lucks, K.E., Schindler, D. (2015) Unleashing Animal Spirits - Self-Control and Overpricing in Experimental Asset Markets. Working paper.

#066 Kranz, T.T., Teschner, F., Weinhardt, C. (2014) User Heterogeneity in Trading Systems: Assessing Trader's Market Predisposition via Personality Questionnaires. Proceedings of the 2014 47th Hawaii International Conference on System Sciences (HICSS) 1: 1230 – 1239.

#067 Kuhn, M., Kuhn, P., Villeval, M.C. (2014) Self Control and Intertemporal Choice: Evidence from Glucose and Depletion Interventions. CESIFO Working Paper No. 4609.

#068 Li, C. (2015) Are the Poor Worse at Dealing with Ambiguity: Comparison of Ambiguity Attitudes between Urban and Rural Chinese Adolescents. Working Paper.

#069 Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., Pardo, S. T. (2012) Individual Differences in Numeracy and Cognitive Reflection, with Implications for Biases and Fallacies in Probability Judgment. Journal of Behavioral Decision Making, 25(4): 361–381.

#070 Lohse, J. (2014) Smart or Selfish - When smart guys finish nice. AWI Discussion Paper Series 578.

#071 Lubian, D., Untertrifaller, A. (2014) Cognitive abilities, stereotypes and gender segregation in the workplace. Economics Bulletin, 32(2): 1268-1282.

#072 Moritz, B., Hill, A.V., Donohue, K. (2013) Individual Differences in the Newsvendor Problem: Behavior and Cognitive Reflection. Journal of Operations Management, 31(1-2): 72-85.

#073 Moritz, B., Siemsen, E., Kremer, M. (2014) Judgmental Forecasting: Cognitive Reflection and Decision Speed. Production and Operations Management, 23(7): 1146-1160.

#074 Narayanan, A., Moritz, B. (2015) Decision Making and Cognition in a Multi-Echelon Supply Chain: An Experimental Study. Forthcoming in Production and Operations Management.

#075 Niessen, A.S.M. (2015) University of Groningen Working paper

#076 Nieuwenstein, M.R., Van Rijn, H. (2012) The unconscious thought advantage: Further replication failures from a search for confirmatory evidence. Judgment and Decision Making, 7(6): 779-798.

#077 Noussair, C.N., Trautmann, S.T., van de Kuilen G. (2014) Higher Order Risk Attitudes, Demographics, and Financial Decisions. Review of Economic Studies, 81 (1): 325-355.

#078 Obrecht, N. A., Chapman, G. B., & Gelman, R. (2009). An encounter frequency account of how experience affects likelihood estimation. Memory & Cognition, 37, 632-643. doi: 10.3758/MC.37.5.632.

#079 Obrecht, N. A., Chapman, G. B., Gelman, R. (2007) Intuitive t-tests: Lay use of statistical information. Psychonomic Bulletin & Review, 14: 1147-1152.

#080 Oechssler, J., Roider, A., Schmitz, P. (2009) Cognitive Abilities and Behavioral Biases. Journal of Economic Behavior and Organization, 72(1): 147-152.

#081 Östling, R., Wang, J.T., Chou, E.Y., Camerer, C.F. (2011) Testing Game Theory in the Field: Swedish LUPI Lottery Games. American Economic Journal: Microeconomics, 3(3): 1-33.

#082 Pennycook, G., Cheyne, J.A., Barr, N., Koehler, D.J., Fugelsang, J.A. (2014) Cognitive style and religiosity: The role of conflict detection. Memory & Cognition, 42(1): 1-10.

#083 Pennycook, G., Cheyne, J.A., Barr, N., Koehler, D.J., Fugelsang, J.A. (2014) The role of analytic thinking in moral judgments and values. Thinking & Reasoning, 20(2): 188-214.

#084 Pennycook, G., Cheyne, J.A., Koehler, D.J., Fugelsang, J.A. (2015). Is the cognitive reflection test a measure of both reflection and intuition? Forthcoming in Behavior Research Methods.

#085 Pennycook, G., Cheyne, J.A., Seli, P., Koehler, D.J., Fugelsang, J.A. (2012) Analytic cognitive style predicts religious and paranormal belief. Cognition, 213: 335-346.

#086 Peysakhovich, A., Rand, D.G. (2015) Habits of Virtue: Creating Norms of Cooperation and Defection in the Laboratory. Forthcoming in Management Science.

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

#088 Razen, M., Kirchler, M., Palan, S. (2014) Correlated Information in Markets. Working paper.

#089 Rhodes, R. E., Rodriguez, F., Shah, P. (2014) Explaining the alluring influence of neuroscience information on scientific reasoning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40(5): 1432-1440.

#090 Royzman, E.B., Landy, J.F., Goodwin, G.P. (2014) Are good reasoners more incest-friendly? Trait cognitive reflection predicts selective moralization in a sample of American adults. Judgment and Decision Making, 9(3): 175-190.

#091 Royzman, E.B., Landy, J.F., Leeman, R.F. (2015). Are thoughtful people more utilitarian? CRT as a unique predictor of moral minimalism in the dilemmatic context. Cognitive Science, 39(2): 325-52.

#092 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. Study 2

#093 Shtulman, A., McCallum, K. (2014) Cognitive reflection predicts science understanding. Proceedings of the 36th Annual Conference of the Cognitive Science Society, 2937-2942.

#094 Simonson, I., Sela, A. (2011) On the Heritability of Consumer Decision Making: An Exploratory Approach for Studying Genetic Effects on Judgment and Choice. Journal of Consumer Research, 37(6): 951-966.

#095 Sirota, M., Juanchich, M. (2011) Role of numeracy and cognitive reflection in Bayesian reasoning with natural frequencies. Studia Psychologica, 53(2): 151-161.

#096 Sirota, M., Juanchich, M., Hagmayer, Y. (2014) Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. Psychonomic Bulletin & Review, 21(1): 198-204.

#097 Sirota, M., Juanchich, M., Kostopoulou, O., Hanak, R. (2014) Decisive evidence on a smaller-than-you-think phenomenon: Revising the “1-in-X” effect on subjective medical probabilities. Medical Decision Making, 34(4): 419-429.

#098 Sulitzeanu-Kenan, R., Halperin, E. (2013) Making a Difference: Political Efficacy and Policy Preference Construction, British Journal of Political Science, 43(2): 295-322.

#099 Sun, H., Bigoni, M. (2015) A Good Rule in the Hobbesian Jungle? An Experiment of the Endogenous Adoption of a Social Norm of Trustworthiness. Forthcoming in Frontiers in Social Psychology.

#100 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: 1275-1289.

#101 Toplak, M.E., West, R.F., Stanovich, K.E. (2013) Assessing miserly information processing: An expansion of the Cognitive Reflection Test. Thinking & Reasoning, 20: 147-168.

#102 Trémolière, B., De Neys, W., Bonnefon, J. F. (2014) The Grim Reasoner: Analytical Reasoning under Mortality Salience. Thinking & Reasoning, 20(3): 333-351.

#103 Trippas, D., Pennycook, G., Verde, M.F., Handley, S.J. (2015) Better but still biased: The link between analytic cognitive style and belief bias. Forthcoming in Thinking & Reasoning.

#104 Weele, J.J. van der, Grossman, Z., Andrijevik, A. (2014) A Test of Dual-Process Reasoning in Charitable Giving. UCSB Working Paper.

#105 Yahalom, N. & Schul, Y. (2013) How thinking about the other affects our reliance on cognitive feelings of ease and effort: Immediate discounting and delayed utilization. Social Cognition, 31: 31-56.

#106 Brown, A.L., Viriyavipart, A., & Xiaoyuan, W (2018) Search deterrence in experimental consumer goods markets, European Economic Review, 104: 167-84.

#107 Gideon, N., Nadler, A., Zava, D. & and Camerer, C. (2017) Single-dose testosterone administration impairs cognitive reflection, Psychological Science, 28: 1398-1407.

#108 Hanaki N., Jacquemet N., Luchini S. & Zylbersztejn A. (2016) Fluid intelligence and cognitive reflection in a strategic environment: evidence from dominance-solvable games, Frontiers in Psychology, 7: 1188.

#109 Königsheim, C., Lukas, M., & Nöth, M. (2018) Individual preferences and the exponential growth bias, Journal of Economic Behavior & Organization, 145: 352-69.

#110 Lukas, M. & Nöth, M. (2019) Interest Rate Changes and Borrower Search Behavior, Journal of Economic Behavior & Organization, 163: 172-189.

#111 Neyse L., Bosworth S., Ring P. & Schmidt, U. (2016) Overconfidence, Incentives and Digit Ratio, Scientific Reports, 6: 23294.

#112 Ring P., Neyse L., David-Barett T. & Schmidt U. (2016) Gender Differences in Performance Predictions: Evidence from the Cognitive Reflection Test, Frontiers in Psychology, 7: 1680.

#113 Ruffle, B.J. & Wilson, A.E. (2018) The truth about tattoos, Economics Letters, 172: 143-147.

#114 Sheremeta, R. M. (2018) Impulsive behavior in competition: Testing theories of overbidding in rent-seeking contests, ESI Working paper.