2015

Cognitive Reflection Test: Whom, How, When

Pablo Brañas-Garza
Middlesex University London

Praveen Kujal
Middlesex University London

Balint Lenkei
Middlesex University London

Cognitive Reflection Test: Whom, How, When

Comments

Working Paper 15-25

Cognitive Reflection Test: Whom, how, when[1]

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

Middlesex University London

Abstract

We report the results of a meta-study of 118 Cognitive Reflection Test studies comprising of 44,558 participants across 21 countries. There is a negative correlation between being female and the overall, and individual, correct answers to CRT questions. Taking the test at the end of an experiment negatively impacts performance. Monetary incentives do not impact performance. Overall students perform better compared to non-student samples. Exposure to CRT over the years may impact outcomes, however, the effect is driven by online studies. 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, Glucose and Cognition.

[1] We thank all authors that shared their data with us. We are especially thankful to Tere Garcia, Antonio Espin and Georgui Kolev for their comments, and the audience at the London Experimental Workshop (RHU 2015), Middlesex BBS (2015) and SABE (Sibiu 2015).

[2] Corresponding author: Economics Department, Middlesex University London, The Burroughs, London, NW4 4BT, U.K. branasgarza@gmail.com.

1. Motivation

The Cognitive Reflection Test (CRT) was first proposed by Frederick (2005) and since then has been extensively used in the Experimental Economics and Psychology literature. Frederick proposed the CRT 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.

Frederick (2005) showed 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, NFC, SAT and WPT) and that the test has a (male) gender bias. 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. 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).[3]

Since Frederick (2005), several researchers have adopted the CRT as a measure of cognitive abilities and used it to study its predictive power in decision making (e.g. Oeschler et al 2009, Campitelli and Labollita 2010, Hoppe and Kusterer 2011, Besedes et al 2012, Andersson et al 2013, Moritz et al 2013 etc.). Oechssler et al (2009) investigate whether behavioral biases are related to cognitive abilities. Replicating the results of 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 2014)

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 2015 etc.). It is a useful test to measure strategic behavior as it not only captures the reflective processes but also the ability to execute small computational tasks (Corgnet et al 2015). For example, Brañas-Garza et al (2012) investigate the relationship between CRT outcomes and subjects' performance in the repeated feedback-free p-Beauty Contest Game (BCG) (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. It seems that the CRT helps us in identifying sophisticated subjects who play according to the Nash equilibrium in this strategic environment. Kiss et al (2015) look at the effect of CRT on withdrawal decisions in an extended version of Diamond and Dybvig's (1983) bank-run game.

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

They find that participants with higher cognitive abilities (as measured by the CRT) tend to identify the dominant strategy easier when strategic uncertainty is present in the game.

It is now well established in the Experimental Economics and Psychological literature that subjects with better 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, Cueva-Herrero et al 2015, Peysakhovic and Rand 2015, Ponti and Rodriguez-Lara 2015). 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) show that if subjects with only low cognitive abilities are trading in an experimental asset market it will lead to bubble formation. While, in markets with only highly cognitive individuals assets 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.

Males generally score significantly higher on the CRT than females (e.g. Frederick 2005, Hoppe and Kusterer 2011, Cueva-Herrero et al 2015, Holt et al 2015 etc.). It has been well documented in the experimental literature that in general 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.). We test for whether the hypothesis regarding the reported gender differences holds in a large sample comprising of very different studies (e.g. different locations, lab based, incentivized, non-student samples etc.).

An important question both in economics and psychology has been regarding the use of incentives in experiments. The available evidence supports both viewpoints suggesting that whether incentives matter or not may be context dependent. Riedel et al (1988) and Scott et al (1988) find a positive relationship between monetary incentives and performance levels while 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)[4]. In this paper we aim to test whether monetary incentives significantly impact the number of correct responses in the CRT.

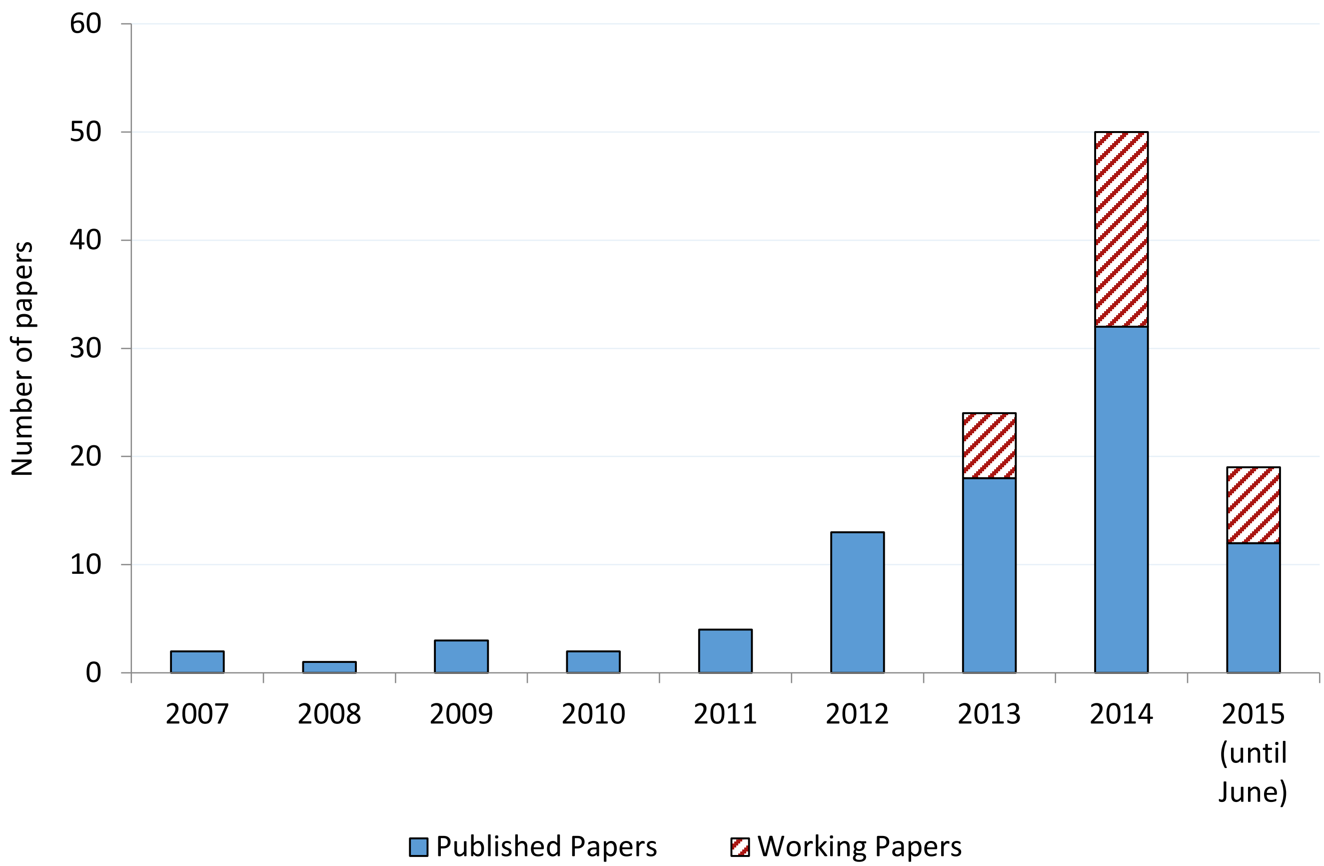
The issue of external validity is important for the experimental literature. There are mixed views on whether studies conducted with (volunteering) university students provide reliable results (Peterson 2001, Levitt and List 2007, Falk and Heckman 2009, Falk et al 2013, Exadaktylos et

[4] 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.

al 2013). Common objections are that student subject pool sample sizes are small and not representative. Our meta-study also compares the CRT results for student and non-student samples of participants.

Figure 1 presents the total number of working and published papers included in our analysis over the period of 2007 to 2015. One can see that in recent years the CRT has been increasingly used. It has been argued that due to its increasing popularity subjects may have had prior experience with the test (Toplak et al 2014).

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



The emergence of popular online experimental platforms such as the Amazon Mechanical Turk (AMT henceforth) (for review on AMT see Paolacci et al 2010, Buhrmester et al 2011, Goodman et al 2013) may have contributed to the acceleration of this process. In later sections we discuss whether the year a particular CRT study was conducted and whether it was conducted on line affected test scores. This is closely related to another issue that is important in the experimental literature, that is, 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.). In our meta-analysis we also address whether the use of computerized settings matter.

The paper is organized as follows. Section 2 presents the procedures and techniques used for data collection. Section 3 provides the main findings. Section 4 explores and discusses these in further detail and provides an overview. Section 5 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. 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 fraction of females among the total).

- How many of the total answered the *B&B*, *Machines*, and *Lillypad* questions correctly (and the fraction of females among them).
- Out of the total how many participants answered all *Three*, *Two* or *One* question(s) correctly (and the fraction 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.

Figure A1 (Appendix) presents a screen shot of the actual questionnaire that researchers were asked to fill out.

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 Behavioural Economics, Management and Psychology with researchers from 21 different countries[5]. 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. The full sample of 44,558 subjects was broken down into further sub-categories. These were:

- *Female* (vs Male=0).
- *Computerized* (vs paper and pencil=0).

[5] 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, USA.

- *Students* (vs Non-students=0).
- Position (whether the CRT was conducted *before*, *in-between* or *after* experiments).
- *Sequence* (the order in which the CRT questions were asked).
- Monetary *incentives* (whether the experimenter paid monetary incentives for correct answers).
- *Visibility* (the year in which the studies were conducted, see also Table 2).

Appendix Table A1 includes a breakdown regarding the number of observations available in each category in our sample.

2.3. Empirical strategy

We use OLS[6] regressions to estimate the relationship between CRT outcomes and the list of variables defined earlier. The robust standard errors are clustered around study IDs. Note that our meta-analysis includes 118 studies and there is a remarkable heterogeneity among them (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 1) with six additional sub-samples (see Appendix):

- A sub-sample including female subjects only (Appendix Table A2). In section 4.1 we analyze the impact of gender differences on CRT results.
- A sub-sample excluding studies where participants were university students (Appendix Table A3). In section 4.3 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 A4). In section 4.4 we analyze the impact of positioning of the CRT compared to

[6] Other statistical models such as probit and logit provide similar results.

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.

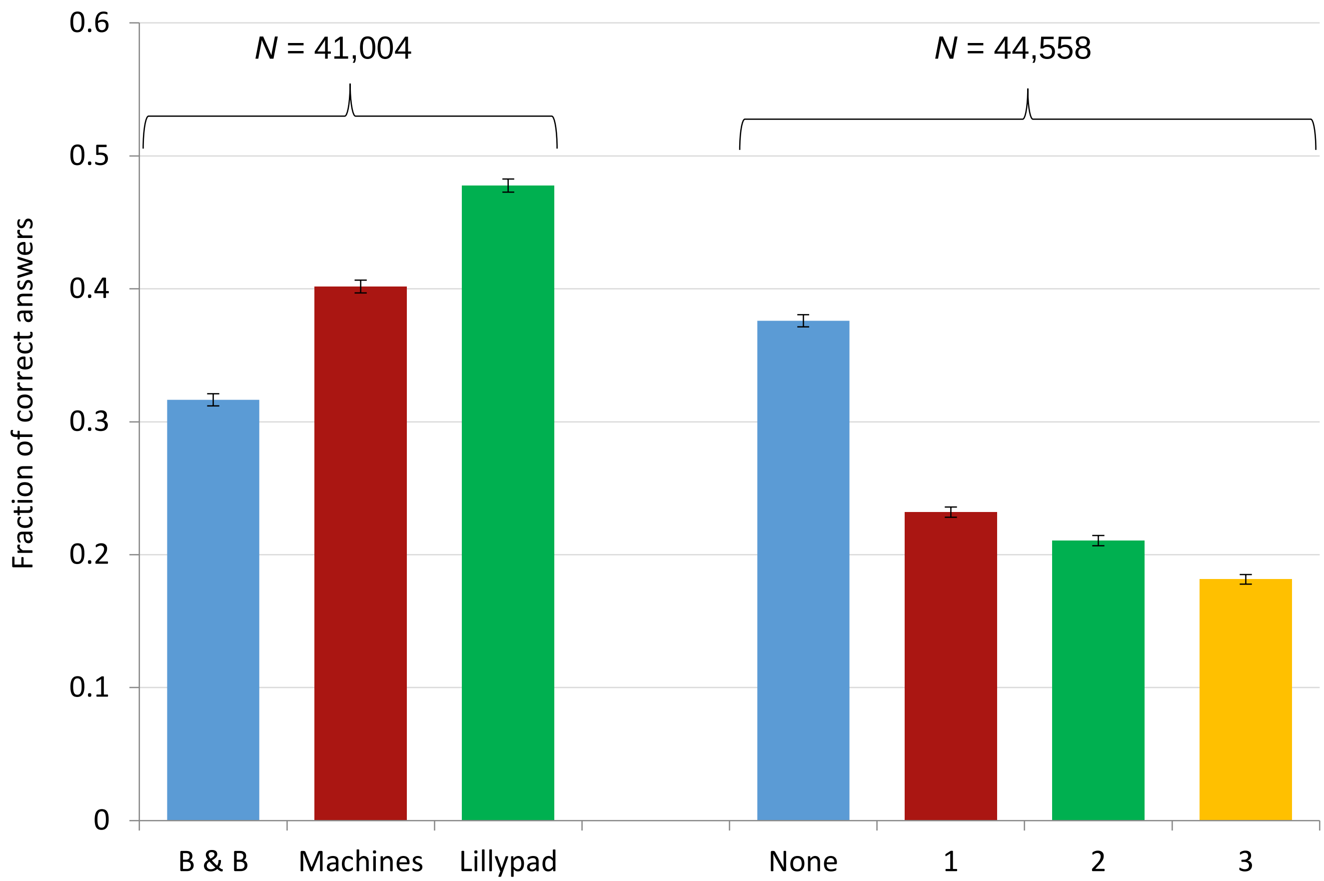
- A sub-sample excluding the studies where the sequence of the questions were randomly determined (Appendix Table A5). In section 4.5 we analyze the effect of the CRT question sequences on test outcomes. We divide our full sample between standard sequence (i.e. *B&B, Machines, Lillypad*) and other sequences. 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.
- A sub-sample excluding studies where monetary incentives were used to reward correct answers (Appendix Table A6). In section 4.6 we analyze the impact of monetary incentives on CRT performance.
- A sub-sample excluding studies where the experimenters used Amazon Mechanical Turk (Appendix Table A7). In section 4.7 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[7].

3. Basic findings

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

Figure 2 shows a summary of our results for the correct answers by individual questions and for the test as a whole.

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



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 seems that the *B&B* question is more cognitively demanding for the subjects. 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<0.001)*.

The right hand side of Figure 2 gives the results based upon the total number of correct answers, i.e. *None, 1, 2, 3* (*N* = 44,558)[8]. Note that 38% of the participants provide none meanwhile, 18% provide all correct answers. Our results indicates that a third of the population lack reflective, or cognitive, abilities. Meanwhile, the remaining 62% have at least some. As previously, the 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 *($p<0.01$)*.

4. Whom, how, when

4.1. Gender bias

Frederick (2005) (*N* = 3,428) showed that males perform better in the CRT (also see Oechssler et al. 2009, Hoppe and Kusterer 2011, Cueva-Herrero et al 2015, Holt et al 2015, etc.). We obtain similar results (*N* = 44,558; females 52.76%) (Figure 3). 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 even when we control for test characteristics (e.g. monetary incentives, computerized, student samples, positioning of the experiment etc.) (see row 1, Table 1).

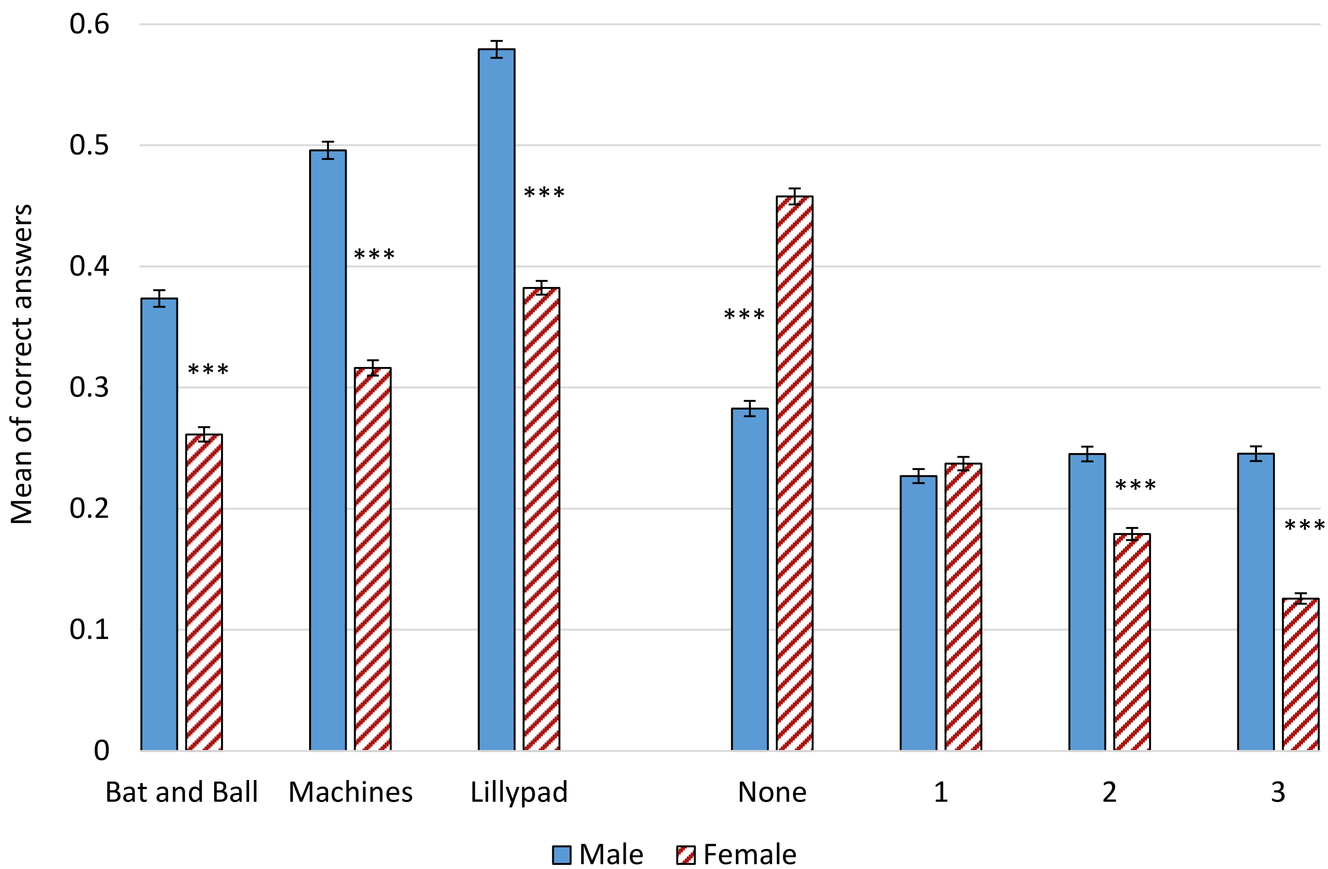
--- Insert Table 1 here ---

Knowing that the CRT has a strong male bias is useful 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

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

females. This implies that we not only select highly cognitive individuals, but also that the sample is strongly biased towards males.

Figure 3: Mean of correct answers by gender.



Note: The asterisks reflect the p-values from the regressions analysis (Table 1, row 1).

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 however 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.

Tables A3, A4, A5, A6, A7 (Appendix) report the results from various robustness checks conducted to test the validity of our model (see section 2.3). The initial results on gender differences (Table 1) remain negative and statistically significant ($p<0.01$) throughout. In addition we replicated the regressions with a female only sample (Table A2). 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.

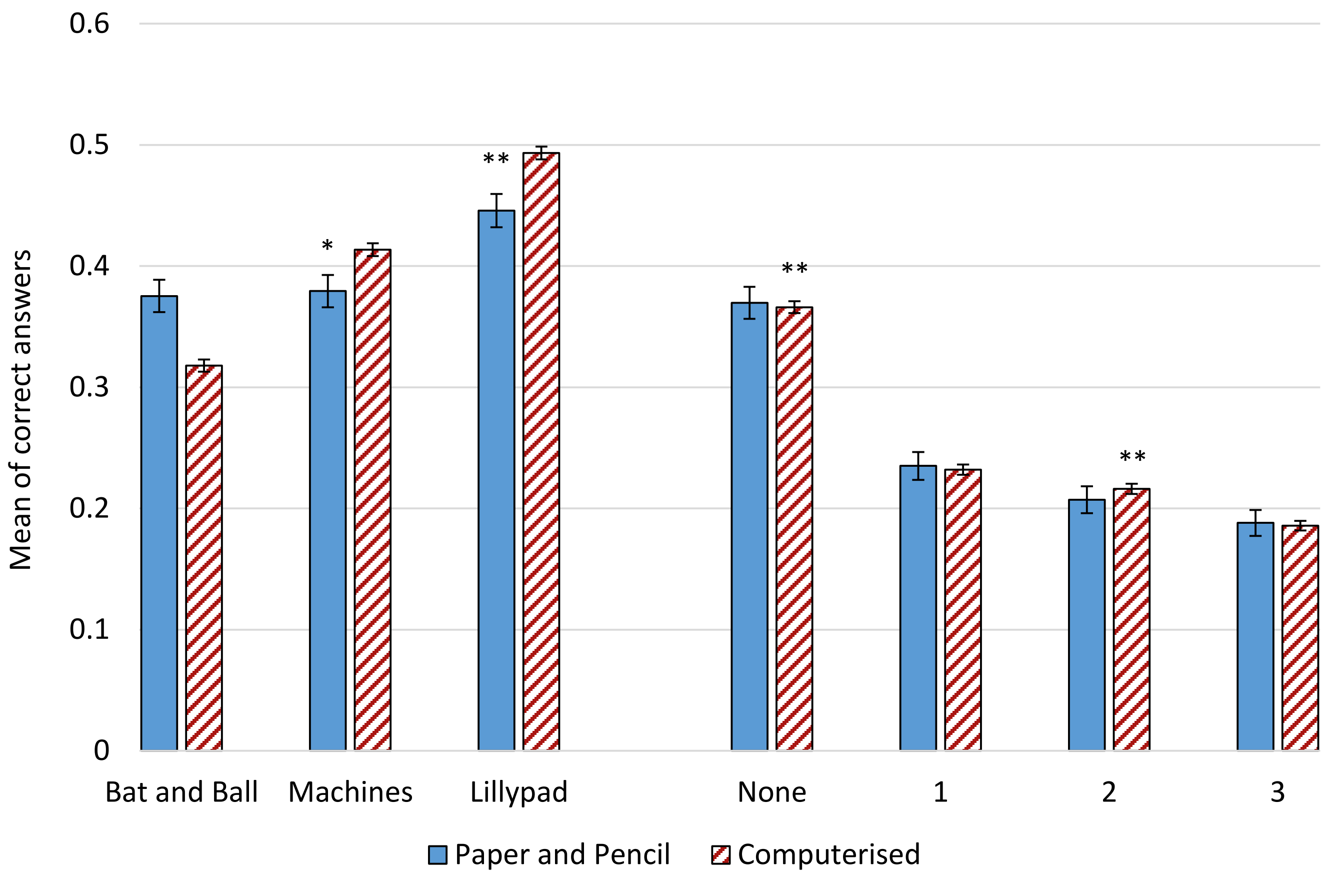
4.2. Hand run vs. computerized?

Figure 4 presents the mean of correct answers for the CRT questions for hand run -paper and pencil- vs. *computerized* studies (12.09% and 87.91% of the full sample, respectively). The regression results in Table 1 (row 2) find that the dummy variable for *computerized* is only weakly significant. It seems that 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 *($p>0.1$)*. However, 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 *($p<0.05$)*. We find this puzzling since one would expect that using paper and pencil would be more conducive to obtaining correct answers.

The robustness checks in the Appendix corroborate these findings. While Tables A2, A4, A6 report somewhat stronger effects, Tables A3, A5 and A7 report identical 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 effects on test scores.

Figure 4: Mean of correct answers by computerized/paper and pencil test.

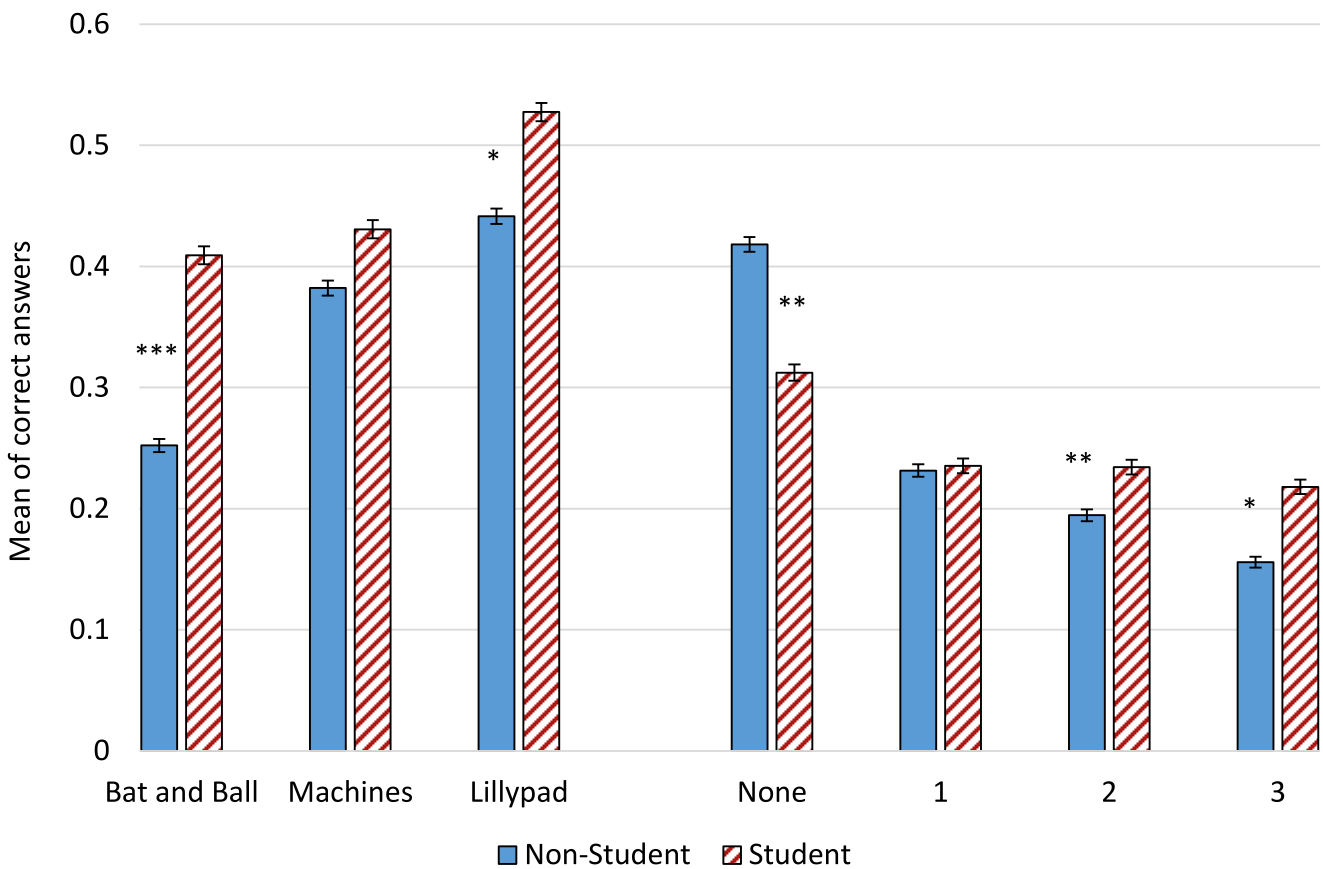


Note: The asterisks reflect the p-values from the regressions analysis (Table 1, row 2).

4.3. Students vs. non-students

Most economics experiments are run with university students. This has raised an obvious question about external validity of experimental data. In recent years there has been a number of papers analyzing this (e.g. Levitt and List 2007, Falk and Heckman 2009, Exadaktylos et al 2013). Our sample includes several studies that were conducted with university students (42.28% of all observations) and others with non-student samples. We find that students score significantly better in the *B&B* and, only slightly better in the *Machines* and *Lillypad* question (Figure 5). The right hand side of Figure 5 shows 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.

Figure 5: Mean of correct answers by student.



Note: The asterisks reflect the p-values from the regressions analysis (Table 1, row 3).

Table 1 (row 3) confirms the findings in Figure 5. 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.

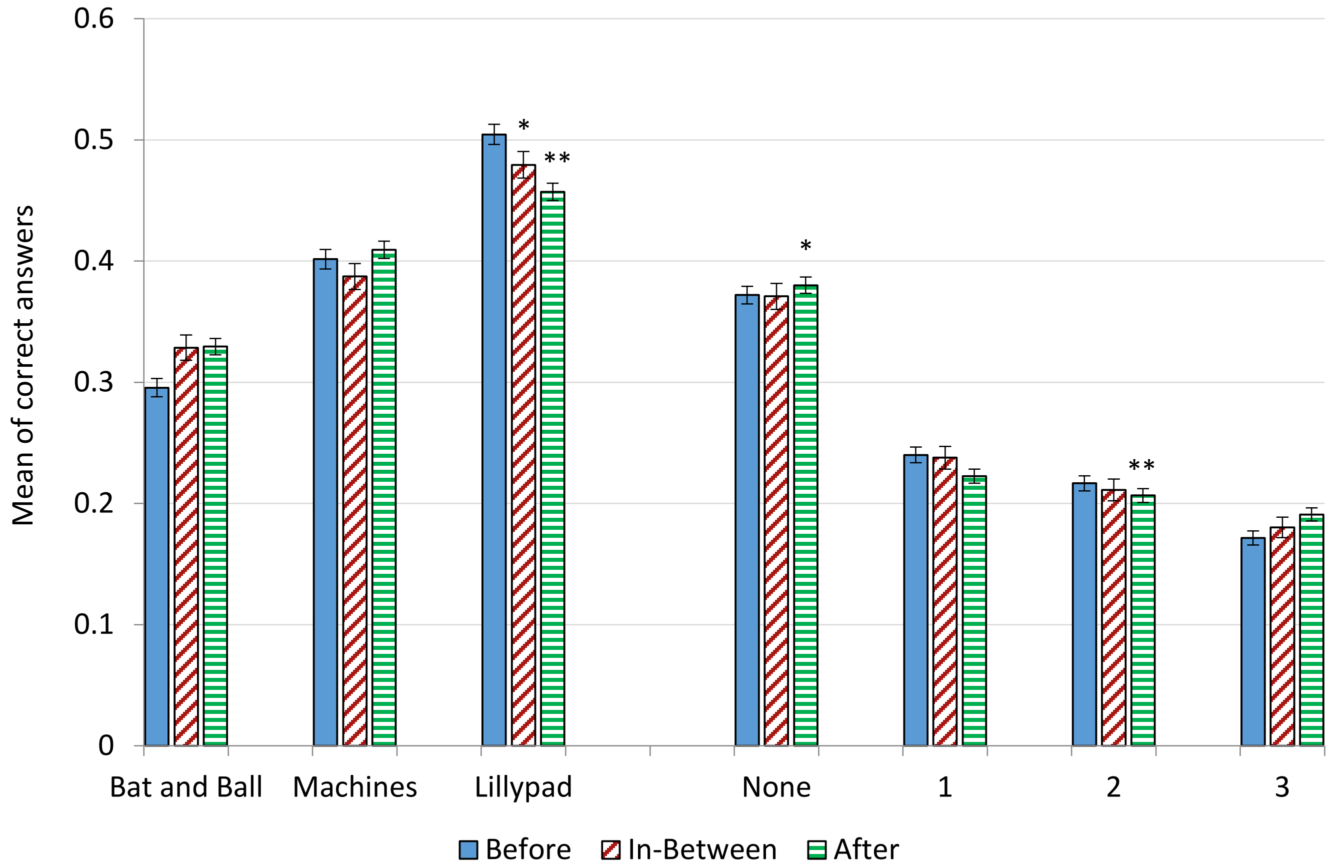
We performed a robustness check (Table A3) only including non-student samples. However, we do not find large differences between Table 1 and A3. The gender bias is identical for both students and non-students samples. The robustness check in Table A2 shows that these effects are stronger when using a female only sample. Tables A4, A5, A6, A7 report results with similar signs but with less statistical power. In sum, our results allow us to state that one can expect the average CRT scores to be higher when using *student* samples.

4.4 When?

It is important to understand whether the implementation of the test *before*, *in-between* or *after* the experiment (*37.66%, 17.75%* and *44.58%* of our sample) affects outcomes. A priori one would expect no differences. However, it has been shown that brain activity is reliant on blood glucose levels as it affects the firing of neurons (Weiss, 1986). Experimental tasks require 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.

Looking at Figure 6 (rows 4a and 4b in Table 1) one sees that there are significant differences in CRT performance depending upon whether it was conducted *before*, *in-between* or *after* the experiment. 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 1). 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 remain negative throughout (row 4b, Table 1). This suggests that conducting the CRTs *after the experiments* can potentially impact outcomes negatively.

Figure 6: Mean of correct answers by the position of the test compared to the main experiment.



Note: The asterisks reflect the p-values from the regressions analysis (Table 1, rows 4a and 4b).

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 A4, 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 A4).

The robustness checks in Tables A2, A3, A5, A7 report results with similar signs but with identical or somewhat less statistical power compared to the main results in Table1. In contrast, A6 reports stronger effects.

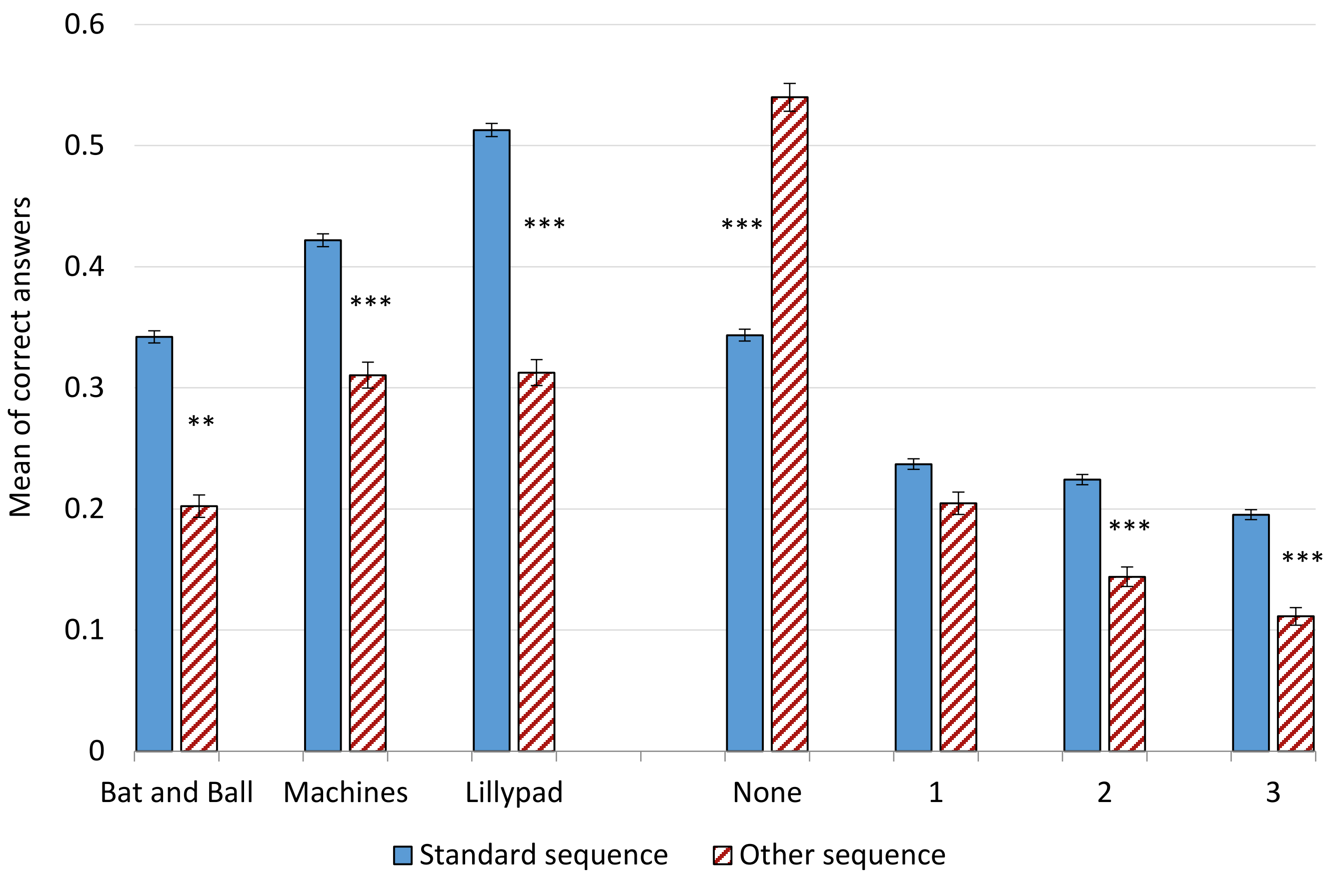
These results are important considering the argument that glucose levels in the brain play an important role in cognition. Effortful, controlled or executive processes and tasks (e.g. experiments) require more glucose than simpler, less effortful or automatic processes. When glucose levels are low, cerebral 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 CRT results.

4.5 The sequence of questions

Frederick (2005) proposed the CRT questions in the following order: *B&B*, *Machines*, and *Lillypad*. This has become the most commonly used sequence and *83.78%* of our sample corresponds to this. Figure 7 indicates that subjects score better on the CRT when the questions are presented in the standard order. Row 5 in Table 1 is consistent with the findings in Figure 7 as the coefficient on *standard* sequence is highly significant for the *B&B* ($p<0.05$), *Machines* ($p<0.01$) and *Lillypad* ($p<0.01$) questions. Looking at the results in row 5 (Table 1) one can also conclude that 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.

However, one may argue that our control group other sequences includes studies where the order of the questions was randomized (11.64% of all of our observations). The randomized sequences can also include questions asked in a standard way with probability 1 in 6. Therefore, in Table A5 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 A5). We cannot thus conclude that the standardized sequence would bias responses in the CRT.

Figure 7: Mean of correct answers by the sequence of the questions.



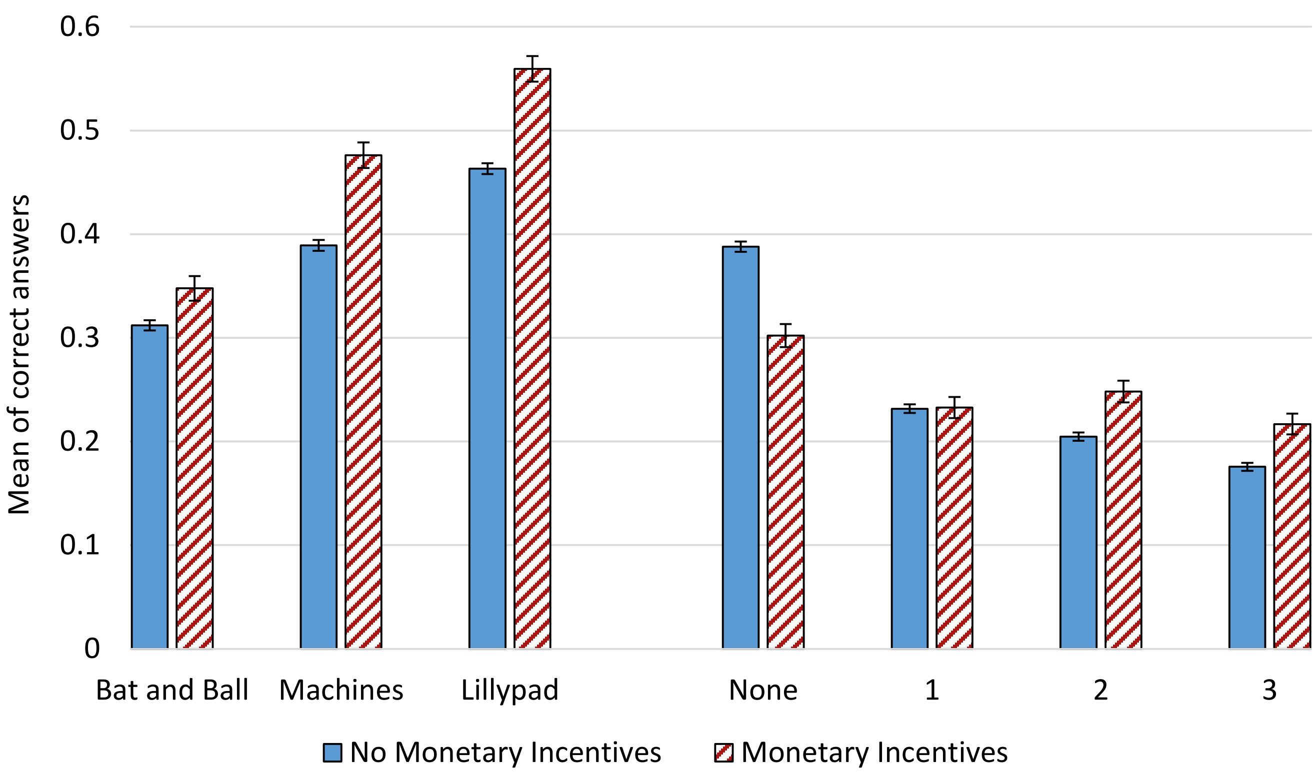
Note: The asterisks reflect the p-values from the regressions analysis (Table 1, row 5).

4.6. Incentives

The effect of financial incentives on human behavior has been a long debated issue in the economics and psychology literature (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 6, Table 1) 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.

Figure 8: Mean of correct answers by monetary incentives.



Note: No asterisks are shown since there are not significant results in the regression (Table 1, row 6).

The series of robustness checks using different samples in Tables A2, A3, A4, A5, A7 (Appendix) for our regressions seem to indicate the same: While three tables (A2, A3 and A4) show a marginal effect, in Tables A5 and A7 the no-effect of *monetary incentives* remains persistent throughout. In addition, Table A6 presents the regression results excluding studies that use monetary incentives to reward correct answers. However, the overall results in Table A6 do not seem to contradict previous findings in Table 1.

The role of incentives with regard to the degree of cognition can also be important (Awasthi and Pratt, 1990). They find that the effectiveness of monetary incentives depends on the cognitive skill of the decision maker. That is, monetary incentives were associated with higher performance only for high cognition individuals. We cannot comment on whether there is a relation between cognition and incentives. Note that in our data both the measure of IQ and performance is the same variable, i.e. the number of correct answers. 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 lack specific details on how and the quantity subjects were paid.

4.7. Exposure to the CRT over the years (visibility)

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. Another issue with testing this conjecture is that some studies are conducted on-line. Answers to the CRT are easily available on line and this sheds doubt on its studying its efficacy using on-line studies. We investigate these issues below.

Table 2 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 1 for studies conducted in 2007, 2 for 2008 and so on.

--- Insert Table 2 here ---

In Table 1 (row 8) 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 8, Table1) 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 (see Table A2, A3, A4, A5 and A6) provide similar weak findings for the exposure conjecture.

However our results dramatically change when AMT studies – where participants have immediate access to the internet- are excluded from the sample. We replicated the regressions excluding all AMT studies (row 8, Table A7). We now 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 success.

5. Discussion

The CRT has become increasingly popular in predicting reflection in economic and psychology experiments. We conduct a meta-survey of the methods employed in 118 studies (*N* = 44,558) across several countries. Our goal was to study whether different forms of implementation mattered in terms of performance on the CRT and consequently the predictive power of the test in those studies. We have three important results.

First, we find that the gender bias result first reported in Frederick (2005) is robust. Men perform notably better in this test. If one is interested in constructing samples based on cognitive ability then this could lead to (gender) sample imbalance. For instance, if one uses three correct answers then the sample is disproportionately biased towards males.

Second, we find statistical evidence to support the argument that *monetary incentives* do not play an important role in improving CRT performance. Note however that we do not have data on the amount, or how, subjects were paid. Due to this the extent of our result is limited.

Third, we find that conducting the CRT after the experiments negatively effects test outcomes. Conducting the test after decreases the probability of obtaining correct answers, meanwhile, the probability of obtaining *None* is increased. This result is important as it provides an indirect support to the argument that glucose is important in cognitive tasks and cognition declines with time and effort. After removing studies from the data where the researchers did not run experiments we find even more significant results.

We also find that *students* are more likely to answer all three questions correctly compared to non-students, and less likely to have zero correct answers. We test for the year effect (*visibility*) and find no clear evidence that exposure positively affects tests results. Regarding the *standard sequence* of the questions, only a marginal effect appears after removing studies where researchers used randomized sequences. Finally, comparing test scores for hand-run vs. *computerized* tests we found a weakly positively significant effect of computerized implementation of the test.

Table 1: Regression analysis

	(1) *B&B*	(2) *Machines*	(3) *Lillypad*	(4) *None*	(5) *1*	(6) *2*	(7) *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) 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)
(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) 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)
(6) 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)
(7) 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)
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-sq	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.

Table 2: 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.

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. (2014) Faith in Intuition and Cognitive Reflection. University of Cologne Working Paper. Study 3.

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.

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. and 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. and 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-Rosa, C., Meissner, T., Bosch-Domenech, A. (2015) Cognitive Bubbles. SFB 649 Discussion Paper 2015-006.

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.

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

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., Espín, A.M., Hernán-González, R. (2015) The cognitive basis of social behavior: cognitive reflection overrides antisocial but not always prosocial motives. ESI Working Paper 15-04.

Corgnet, B., Espin, A., Hernan-Gonzalez, R., Kujal, P., Rassenti, S. (2015) To trust, or not to trust: Cognitive reflection in trust games. Forthcoming in *Journal of Behavioral & Experimental Economics*.

Corgnet, B., Hernan-Gonzalez, R., Kujal, P. and Porter, D. (2014) The Effect of Earned Versus House Money on Price Bubble Formation in Experimental Asset Markets, *Review of Finance*, **1-34.** [doi: 10.1093/rof/rfu031]

Cueva-Herrero, C., Iturbe-Ormaetxe, I., Mata-Prez, E., Ponti, G., Yu, H., Zhukova, V. (2015) Cognitive (Ir)reflection: New Experimental Evidence. Forthcoming in *Journal of Behavioral and Experimental Economics*.

Diamond, D.W., Dybvig, P.H. (1983) Bank Runs, Deposit Insurance, and Liquidity. *Journal of Political Economy*, **91(3):** 401-419.

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. and 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.

Gneezy, U., Rustichini, A. (2000a) Pay enough or don't pay at all. *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.

Holt, C.A, Porzio, M. and Song, M.Y. (2015) Price Bubbles, Expectations, and Gender in Asset Markets: An Experiment. University of Virginia working paper.

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. and 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. (2015) Think Twice Before Running! Bank Runs and Cognitive Abilities. Forthcoming in Journal of Behavioral and Experimental Economics.

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.

Mau, W.C. and 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. *American Economic Review*, **85(5):** 1313-26.

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.

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.

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. Forthcoming in *Management Science*.

Ponti, G., Rodriguez-Lara, I. (2015) Social Preferences and Cognitive Reflection: Evidence from Dictator Game Experiment. Forthcoming in *Frontiers in Behavioral Neuroscience*.

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.

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

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

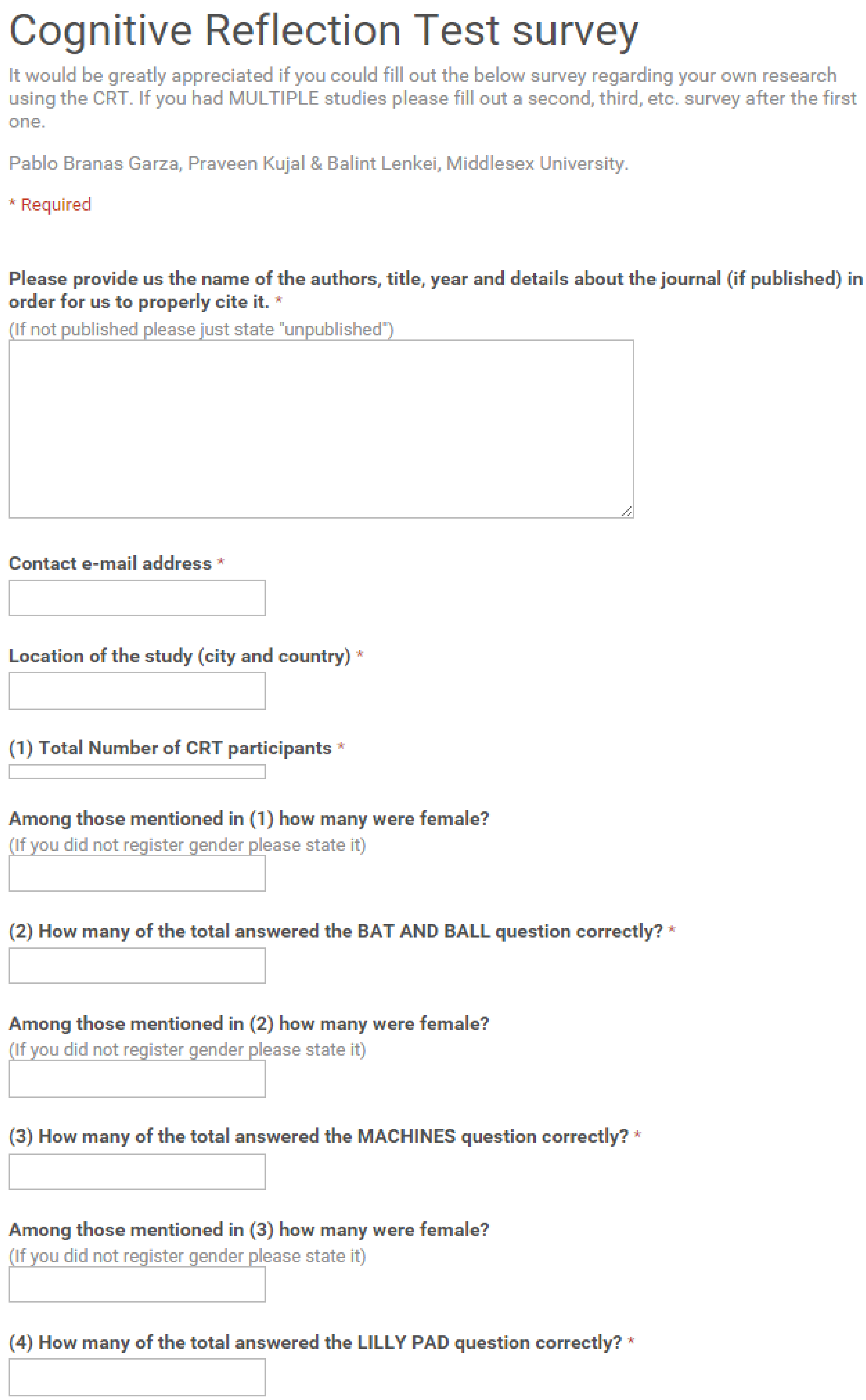
Toplak, M.E., West, R.F., Stanovich, K.E. (2011) The Cognitive Reflection Test as a predictor of performance on heuristics and biases tasks. *Memory & Cognition*, **39:** 1275-1289.

Toplak, M.E., West, R.F. and 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)

Submit

Table A1: 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 (computerized or paper and pencil)	**42,797**	**39,603**
Computerized	87.91%	89.65%
Paper and Pencil	12.09%	10.35%
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 (monetary incentives)	**44,558**	**39,603**
Incentivized	14.67%	15.82%
Non-Incentivized	85.33%	84.18%
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%

Table A2: Robustness check: Females only

	(1) *B&B*	(2) *Machines*	(3) *Lillypad*	(4) *None*	(5) *1*	(6) *2*	(7) *3*
(1) 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)
(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) 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)
(5) 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)
(6) 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)
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-sq	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 A3: 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) 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)
(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) 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)
(5) 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)
(6) 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)
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-sq	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 A4: 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) 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)
(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) 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)
(6) 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)
(7) 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)
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-sq	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 A5: 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-sq	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.

Table A6: Robustness check: excluding studies where the experimenters used monetary incentives to reward correct answers

	(1) *B&B*	(2) *Machines*	(3) *Lillypad*	(4) *None*	(5) *1*	(6) *2*	(7) *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) 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)
(3) 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)
(4a) 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)
(4b) 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)
(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) 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)
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-sq	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 A7: 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) 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)
(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) 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)
(6) 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)
(7) 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)
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-sq	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.

Appendix B

List of articles included in the meta-study

#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.

#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.