ON COGNITIVE ABILITY AND LEARNING IN A BEAUTY CONTEST[1]

Oliver Schnusenberg[2]
Andrés Gallo[3]

Abstract

We reinvestigate a version of the beauty contest originally developed by Keynes (1936) with a focus on cognitive reflection. Using a sample of 166 undergraduate students at a regional university in Florida, we confirm previous research by Burnham et al. (2009) that cognitive reflection, as measured by Frederick's (2005) cognitive reflection test, matters in the first round of the game; players with a higher CRT score pick significantly lower numbers, and their responses cluster more. Unlike previous research, however, we find that cognitive ability is important only when faced with a new situation. In subsequent rounds of the game, cognitive ability is subordinate to a learning effect and players' responses and the variability of responses are not significantly related to CRT scores. This finding is important in financial markets, since it implies that anticipating the decisions and actions of other players is a function of experience, not necessarily cognitive ability.

Key Words: beauty contest, cognitive reflection test, cognitive ability, CRT

JEL Classification: A22

Introduction

The beauty contest was first discussed numerically by Moulin (1986), although it was introduced originally by Keynes (1936). In this contest, players choose a number between 0 and 100. The winning entry is the one closest to a given percentage (most often 2/3) of the mean choice of players. The Nash equilibrium of this game is that everyone chooses 0. Typical experiments, however, show that the typical winning entry is about 17 (see, for example, Thaler 1998, Nagel 1995). Montier (2010) references this game as well to illustrate how difficult it is to incorporate everybody else's decision making process into your own. Also recently, Burnham et al. (2009) utilized a cognitive test to see if the cognitive ability of an individual is associated with the response in the Beauty Contest. They found that higher cognitive ability improves the performance in the Beauty Contest game. This paper complements and expands their research since it uses a cognitive test in the selected group of subjects but it assesses whether the cognitive ability holds in repeated games. This paper shows that, similarly to Burnham et al. (2009), cognitive ability is important in the first round of the game, but once the game is played repeatedly this advantage disappears. This result adds a very important dimension to behavioral economics since it indicates that cognitive ability can be important when dealing with new situations, while repeating interaction creates a feedback mechanism for participants that can reduce the effect of any advantage in cognitive ability.

---

[2] Associate Professor of Finance, Department of Accounting & Finance, University of North Florida
[3] Associate Professor of Economics, Department of Economics & Geography, University of North Florida

The remainder of this paper is organized as follows. The hypotheses are presented in next. The data is presented in the third section, followed by the results. The paper concludes and presents some implications in the last section.

Hypotheses

Individuals tend to think in either an X System or a C System. The X System represents the emotional approach to decision making, while the C System is a more logical way of processing information. The C System is deliberate, deductive, and logical, while the X System is automatic and effortless. Klein (1999) summarizes the conditions under which people are more likely to use the more automatic system[4]:

- When the problem is ill structured and complex
- **When information is incomplete, ambiguous, and changing**
- When the goals are ill-defined, shifting, or competing
- **When stress is high, because either time constrains and/or high stakes are involved**
- **When decisions rely upon an interaction with others**

The highlighted areas are present in the numerical beauty contest. Clearly information is incomplete, since no player knows how the other players will play the game. Moreover, the optimal decision for a player depends on the expectations about the other players' play of the game. In addition, if we view the numerical beauty contest as a corollary to the stock market, playing the stock market game (i.e., guessing the price everybody else is willing to pay for a stock) results in high stress because a trader's job may require swift execution and/or substantial amounts of money.

Given this context, it stands to reason that most people will use System X in order to play the numerical beauty contest. Frederick (2005) developed a Cognitive Reflection Test (CRT) consisting of three questions, which can easily be used to measure the ability of the C-System to control the X-System. Montier (2010) elaborates: "I've found that the number of Frederick's questions that you get correct correlated with your general vulnerability to a whole plethora of other behavioral biases….Those who get zero questions right seem to suffer more pronounced examples of the biases than those who get three questions right." Burnham et al. (2009) used a cognitive ability test developed by the company Assessio, which is consistent with the test that we use in our experiment.

We expect that those contestants that answer more questions correctly on the CRT will play the game differently than those who answer zero or only one question on the test correctly. Specifically, the CRT was designed to assess a specific cognitive ability; it assesses individuals' ability to suppress an intuitive and spontaneous ("system X") wrong answer in favor of a reflective and deliberative ("system C") right answer.[5] In the remaining discussion, we will use the terms "cognitive ability" and "cognitive reflection" alternative to refer to this intended meaning of the CRT test. In the context of the beauty contest, it stands to reason that C-System users will pick numbers that are closer to the Nash equilibrium of zero, since they are likely to reason more through the thinking that the contestants will follow. This hypothesis is also informed by the empirical results provided by Burnham et al. (2009).

---

[4]   Also summarized in Montier (2010).

[5]   Obrecht, Chapman, and Gelman (2007), for example, find that higher CRT scores are correlated with mean differences in comfort with statistical concepts; Pinillos et al (2011) find that taking the CRT and answering questions correctly activates System C processes for subsequent tasks.

H1: C-System users will pick lower numbers than X-System users in the contest.

Furthermore, if System C users think through the reasoning of other players to a greater extent, then it stands to reason that their responses will be closer together. Consequently, we hypothesize that the standard deviation of their responses will be lower:

H2: C-System user responses will be more clustered and exhibit a lower standard deviation.

Given the empirical results regarding the better performance of subjects with higher cognitive ability, we want to explore whether this advantage holds in repeated instances of the game. To do that, we have the same group of subjects play this game two more times.

Accordingly, our third hypothesis is,

H3: C-System user responses will be consistently lower than X-System users responses in further rounds of the game.

Data

To conduct the contest, we utilized students in an introductory business undergraduate course at a regional university in Florida. Students were asked to answer Frederick's (2005) three CRT questions as well as the numerical beauty contest question, which are reproduced in the Appendix. In the first round, 166 students participated in the survey.

After the first round, students were shown the distribution of answers as in Figure 1.

Figure 1. Distribution of Round 1 Answers in the Beauty Contest.
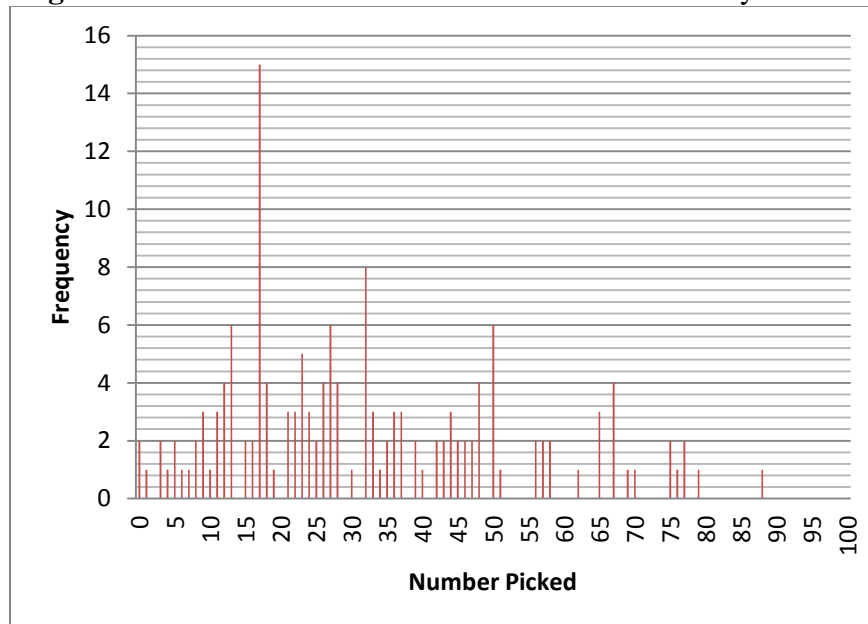
Students were also told the average and winning entry. No additional information was provided.[6] Students were then told that the contest would be played again[7] and, while anyone was welcome to play and anyone could win, the distribution and winning number would be

---

[6]   In this environment we wanted to reflect market situations as accurately as possible. In any normal financial market operation, market participants make their decisions, observe the results, and then use that information as feedback for their decision next time.

[7]   Camerer and Ho (2000) find that the responses converge to the Nash equilibrium of zero after about 10 rounds.

based only on the distribution of those students who participated in the contest the first time.[8] We added this requirement so that the students who participate in both rounds adjust their play of the game only in response to others playing the game repeatedly, not to newcomers to the game who have not had a chance to "learn" how to play the game. 94 students participated in the second round.

Students were then told that the contest would be conducted a third time. Students were presented with the distributions of the first two rounds (Figures 1 and 2) and the averages in each round. Again students were told that anyone could play and win, but the distribution that determines the winner would be based only on those students who played both previous rounds. Seventy students participated in the third round.

Figure 2. Distribution of Round 2 Answers in the Beauty Contest.
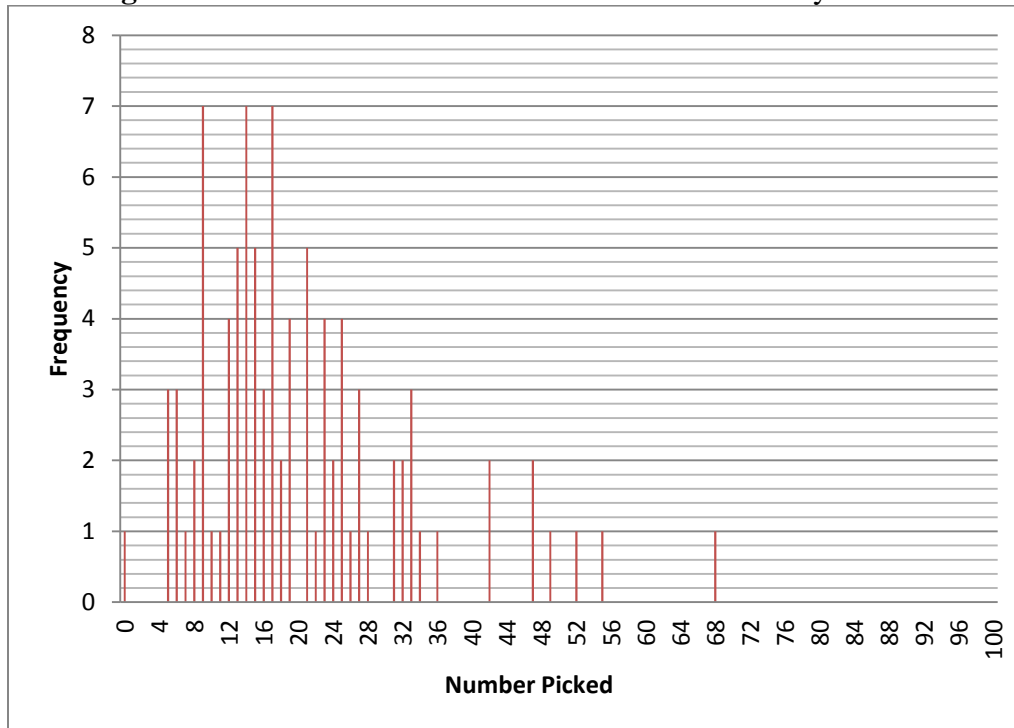
As compensation for participating in the survey, students in each round were offered a quiz question should they win the contest.[9]

---

[8]   We added this requirement to observe how the students that participated in the first round would play the game in subsequent rounds.

[9]   A quiz in the class is worth about 2.5%, with about seven questions per quiz. Simple math shows that this is a very small incentive for participating in the study. Moreover, we did not provide any incentive for answering the CRT questions correctly. This is because we wanted to identify whether students are natural System X or System C thinkers. Our results reveal that those who spend more time on answering the questions on the CRT correctly adjust the answers to the beauty contest more, which is exactly what we predicted. If all participating students simply provided random answers (because the incentive is not high enough), then we would expect the answer to the CRT and the beauty contest to be uncorrelated. We acknowledge that there is a selection bias that may cause random correlation between the beauty contest and the CRT results if only some of the students provide random answers.

Results

The distribution of numbers picked for all three rounds is shown in Figures 1 through 3.

Figure 3. Distribution of Round 3 Answers in the Beauty Contest.
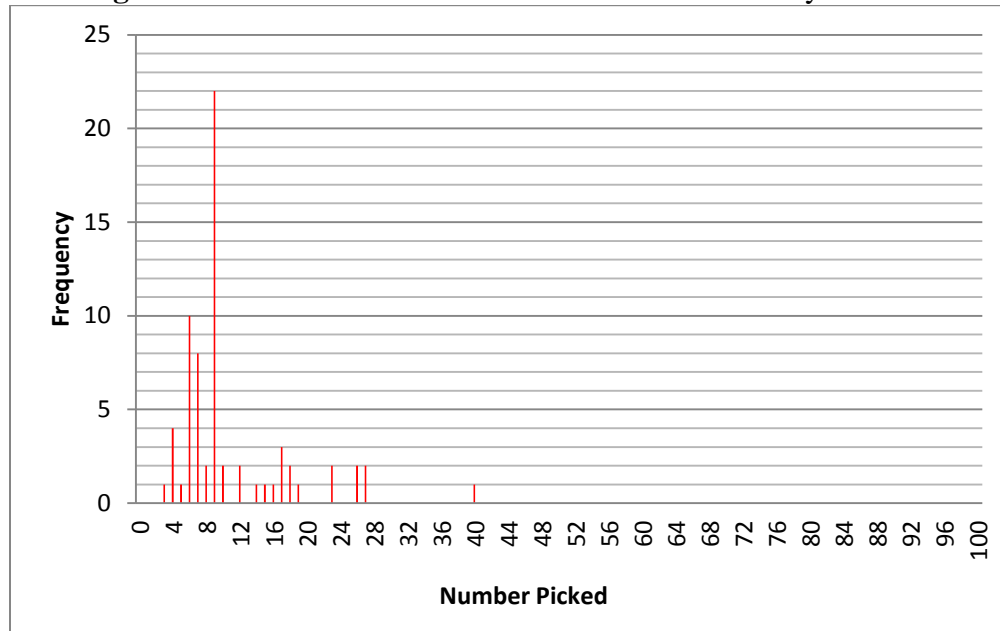
A comparison of the figures clearly indicates that the range of numbers has decreased across the two rounds. In round 1, the average number picked was 32.05, which made the winning entry 21. In round 2, the average number picked was 20.26, which rendered the winning number 14. In the third round, the average number picked was 10.81, with a winning entry of 7^{10}. In round 1, three students picked the number 21, while seven students picked the winning number in round 2 and eight students picked the winning number in the third round. The convergence of the answers toward the Nash equilibrium of zero in later rounds corresponds to the findings of Camerer and Ho (2000).

Descriptive statistics from Frederick's (2005) Cognitive Reflection Test (CRT) are shown in Table 1. Panel A shows the results for the combined results from all three rounds. As is evident from Panel A, the average number picked is lower for students who answer more questions correctly on the CRT. For example, while the average (median) number picked for students who answered no questions on the CRT correctly is 28.90 (23.00) in all three rounds of the beauty contest, the average (median) number picked for those students who answered all questions on the CRT correctly is 21.34 (17.00). The difference in the numbers picked between those who answered none of the questions correctly versus those who answered them all correctly is significant (p-value = 0.015). There is also a decrease in the standard deviation of answers across the categories; for those answering no questions on the CRT correctly, the standard deviation of numbers picked is almost 20, while it is slightly less than 15 for those answering all questions on the CRT correctly. This difference in variances is highly significant (p-value = .005).

---

[10]   All these averages are statistically different from each other. We run a t-test for unpaired samples with different variance and the averages are statistically different at 99.5% of confidence.

Panel B of Table 1 shows the results for round 1 only, Panel C shows the results for round 2 only, and Panel D shows the results for Round 3 only. P-values for tests of differences in means and variances are presented in Panel E. Within each round, the average and median numbers picked decrease consistently as students are more reflective on the CRT test,[11] although the result is much less pronounced in Round 3 of the game. Some of the differences in means in Panel E are statistically significant for different CRT scores in round 1, but most are not significant in rounds 2 and 3.

Table 1. Descriptive Statistics of Frederick's Cognitive Reflection Test in Both Rounds.

Panel A – All Rounds				
Questions Correct	Frequency	Avg. Number Picked	Median Number Picked	Std. Deviation of Answer
0	94	28.90	23.00	19.73
1	64	23.36	20.00	18.62
2	78	22.60	17.00	17.29
3	94	21.34	17.00	14.71
Total	330	24.18	18.00	17.80

Panel B – Round 1				
Questions Correct	Frequency	Avg. Number Picked	Median Number Picked	Std. Deviation of Answer
0	51	38.12	33.30	21.29
1	28	31.48	28.50	19.85
2	41	29.43	28.00	19.84
3	46	28.00	25.75	15.60
Total	166	32.05	28.00	19.53

Panel C – Round 2				
Questions Correct	Frequency	Avg. Number Picked	Median Number Picked	Std. Deviation of Answer
0	25	21.48	21.00	10.43
1	19	23.26	21.00	17.91
2	21	20.10	17.00	8.79
3	29	17.36	16.00	11.16
Total	94	20.26	17.00	12.21

Panel D – Round 3				
Questions Correct	Frequency	Avg. Number Picked	Median Number Picked	Std. Deviation of Answer
0	18	13.07	13.00	6.43
1	17	10.12	8.00	6.13
2	16	8.40	8.50	5.47
3	19	11.32	9.00	8.39
Total	70	10.81	9.00	6.83

---

[11]   The only exception to this is 2 and 3 questions correct for round 3.

Panel E – p-Values for Tests of Different Averages and Variances with Respect to CRT Score

CRT Scores	Round 1 Average	Round 2 Average	Round 3 Average	Round 1 Variance	Round 2 Variance	Round 3 Variance
0 vs 1	0.1707	0.7018	0.1741	0.7075	0.0143**	0.8555
0 vs 2	0.0463**	0.6275	0.0288**	0.6506	0.4414	0.5340
0 vs 3	0.0085***	0.1667	0.4798	0.0363**	0.7427	0.2771
1 vs 2	0.6756	0.4912	0.4005	0.9810	0.0028***	0.6609
1 vs 3	0.4330	0.2101	0.6257	0.1505	0.0239**	0.2129
2 vs 3	0.7114	0.2254	0.3369	0.1185	0.2733	0.0993*

*   Significant at the 10% level
**   Significant at the 5% level
***   Significant at the 1% level

      Similarly, the standard deviation of answers picked is smaller for students who reflect more on the answers they provide in round 1 and, to a lesser extent, round 2. By round 3, the answers from more reflective students tend to have a higher standard deviation.

      The results for round 2 presented in Panel C of Table 1 are less obvious. The average and median numbers decrease for students who scored higher on the CRT, but this trend is less pronounced than for the first round of the contest. The insignificant results of the t-tests, in Panel E of Table 1, corroborate this result. Similarly, while it appears that the standard deviation is decreasing slightly for students with higher CRT scores, the evidence is less convincing in this round; even though it does appear that the variance decreases between students answering one question correctly versus those answering two or three questions correctly, the variance for those students answering one question on the CRT correctly is significantly greater than the variance for those answering none correctly.

      The results for round 3 presented in Panel D of Table 1 are very similar to the round 2 results, with one important difference. Students who scored very high on the CRT appear to both pick higher numbers and exhibit a higher variance than those who scored lower.

      The results presented in Table 1 suggest that the level of reflection that students exhibit is a rather pronounced predictor of student responses in the first round of the contest. This makes sense, as a higher level of reflection or reasoning will lead students to pick a lower number. It also makes sense that this group of students will give responses that cluster more than students who are more impulsive, leading to a lower variance in their responses. As the contest progresses into subsequent rounds, however, it appears that the results on the CRT test become a less useful predictor of student responses as well as the variance of responses. In other words, in later rounds there is a less pronounced difference between those students who think impulsively versus reflectively as measured by the CRT.

      To further investigate whether the CRT results can be used to predict performance in the beauty contest, we next classify students into System X (impulsive) versus System C (reflective) groups. The System X group contains those students with either 0 or 1 answers correct on the CRT, while the System C group contains those students with either 2 or 3 answers correct on the CRT. We then repeat the analysis from Table 1 for these two groups.

The results from the analysis of System X and System C groups are presented in Table 2. As in Table 1, the combined results for rounds 1 and 2 are presented in Panel A, while round 1 and round 2 results are presented in Panels B and C, respectively. As shown in the third line of Panel A, the difference in the average and median number picked between System X and System C users is significant at the 1% and 5% level, respectively. Moreover, the difference in the standard deviation of answers is significant at the 5% level. This indicates that System C users, on average, pick lower numbers and exhibit less variability in their answers than System X users across all three rounds of the contest.

Table 2. System X and C Descriptive Statistics in Both Rounds.

Panel A – All Rounds				
System	Frequency	Avg. Number Picked	Median Number Picked	Std. Deviation of Answer
X	158	26.66	22.50	19.42
C	172	21.91	17.00	15.90
Difference		4.75***	5.50**	3.52**
Panel B – Round 1				
System	Frequency	Avg. Number Picked	Median Number Picked	Std. Deviation of Answer
X	79	35.77	32.00	20.91
C	87	28.67	27.00	17.64
Difference		7.10***	5.00***	3.27
Panel C – Round 2				
System	Frequency	Avg. Number Picked	Median Number Picked	Std. Deviation of Answer
X	44	22.25	21.00	13.97
C	50	18.51	17.00	10.24
Difference		3.74*	4.00*	3.73**
Panel D – Round 3				
System	Frequency	Avg. Number Picked	Median Number Picked	Std. Deviation of Answer
X	35	11.63	9.00	6.37
C	35	9.98	9.00	7.26
Difference		1.65	0.00	-0.89

*　　Significant at the 10% level
**　　Significant at the 5% level
***　　Significant at the 1% level

Panel B shows the results from round 1. For the average number picked and the median number picked, the difference between X and C System users is highly significant, but the difference in the standard deviation of answers is not significant. Thus, while more reflective students apparently pick lower numbers, they do not seem to cluster more. The round 2 results in

Panel C show less convincing results for the average and median number picked. The respective difference between X and C System users of 3.74 and 4.00 is only significant at the 10% level. In round 2, however, the standard deviation of answers is significantly higher for the more impulsive System X users.

Panel C, showing the results from round 3 of the contest, illustrates that the CRT test results become even less important in subsequent rounds. Neither the average numbers picked, the medians, nor the standard deviations differ at conventional significance levels.

The results from Table 1 and 2 combined provide strong evidence that more reflective students pick lower numbers in the beauty contest, particularly in the first round of the contest, than the more impulsive students. Moreover, at the extreme ends of the CRT (0 versus 3 questions correct), there is a significant difference in the variability of answers, with the more reflective students' answers clustering more. While this difference in the standard deviation of responses is not observed in the extreme cases in the second round of the contest, it is observable when System X and System C are classified as (0, 1) and (2, 3) CRT questions correct, respectively. In round 3 of the contest, however, the differences are not significant for either System X or System C users or for the extreme cases.

As an additional test of the relationship between the number picked and the responses on the CRT, we performed a regression analysis. We pool the results from the three rounds and we also present a regression individualizing each round with a different variable. As a result, we ran three simple OLS regressions using the number picked as the dependent variable and the number of correct CRT responses as the independent variable. The regression results are displayed in Table 3.

The columns in Table 3 show the regression results for the three rounds of the beauty contest combined in one variable (Round) and another variable representing the score in the CRT test (CRT Test). Model 1 regresses the number picked on Round and on dummy variables of the CRT score in each round. Model 2 regresses the number picked on Round and on CRT Test. In both models, the results show that the coefficient of the variable Round is negative, which indicates that the score drops strongly in every successive round. Of particular interest is the coefficient of the variable CRT Test in Model 2, which indicates that the number picked decreases by about 2.307 for each additional question answered correctly on the CRT, on average. However, for each additional round played, the average number picked decreases by 10.7 points. This implies that, given that the initial average number was 32.05, it should take about four rounds to reach convergence to zero. This result implies that the learning effect, playing an extra round, is stronger than the cognitive effect, a higher score in the cognitive ability test. This importance of the learning effect is also reflected in the quite substantial adjusted r-squared of .25 for the pooled regression. For each round individually, the adjusted r-squared is a maximum of only .04 for the round 1 regression.

If we run the same regressions but with different variables for the CRT score in each separate round (Model 1), these results are even more telling. The variables labeled CRT*Roundi (i=1,2,3) represents a variable that contains the CRT test scores for each separate round, i. The results show that the CRT test score was highly significant (-3.23) in round 1, marginally significant in round 2 (-1.01) and statistically insignificant in round 3. This implies that subjects with a higher cognitive result will enjoy a strong initial advantage. In the second and third rounds of the game the coefficient for the cognitive test is not as significant, confirming the weaker results for these rounds from Tables 1 and 2. This also reflects the results from the

pooled regression regarding the importance of experience compared to cognitive knowledge in repeated games.

Table 3. Regression Results Using CRT Answers to Predict Beauty Contest Scores (t values in parentheses).

Independent Variables	Model 1	Model 2
CRT*Round1	-3.23***	
	(-3.31)	
CRT*Round2	-1.01**	
	(-2.01)	
CRT*Round3	-0.11	
	(0.22)	
CRT Test		-2.31***
		(-3.20)
Round	-12.87***	-10.73***
	(-7.36)	(-10.04)
Constant	49.62***	46.03***
	(15.25)	(20.18)
	Adj. R2: 0.25	Adj. R2: 0.25
	F-test: 28.64 Pr. = 0.000	F-test: 55.98 Pr. = 0.000
	Number of Observations: 330	Number of Observations: 330

*   Significant at the 10% level
**   Significant at the 5% level
***   Significant at the 1% level

The results indicate that the initial responses students provide may be influenced by the system they use to analyze the information. Moreover, it also appears that the learning that takes place between rounds of the contest may not be influenced by the cognitive system employed by students. Indeed, if we define "learning" as the adjustment in responses between the two rounds, we find that "learning" for System X users is 12.83, while it is 10.64 for System C. The difference of 2.19 is not significant (p-value = 0.29) and neither is the difference in variances in "learning" (p-value = 0.82). The findings for "learning" between rounds 2 and 3 are virtually identical, with a System X user average of 10.31, and a System C user average of 9.17 (p-value of difference test equals 0.34), with an insignificant difference in variances (p-value = 0.77).

Conclusion and Discussion
Overall, it seems that students' initial response is, perhaps, influenced by the cognitive system students use; more reflective students put more thought into the answer they provide and the answer is therefore closer to the Nash equilibrium of zero. That is, System C users' responses are closer to the "textbook" answer and, at least in the extremes, these students' responses cluster more than those of their more impulsive counterparts. It does not appear, however, that either System C or System X users adjust their responses faster, and there is much less convincing evidence that the two groups differ in their responses and the variability of their responses in

subsequent rounds of the contest. In this small setting, adjustment of responses in a beauty contest does not appear to be influenced by cognitive reflection as measured by the CRT.[12]

We confirm previous research by Burnham et al. (2009), which illustrates the importance of cognitive reflection in games like the beauty contest, where the answer depends on the answer other players will provide. Unlike previous literature, however, we find that cognitive reflection is only important in the initial stage of a game, when the players are playing for the first time. This indicates that cognitive ability is important only when faced with a new situation. In subsequent rounds of the game, a player's cognitive ability does not influence the rate of response adjustment.

Our findings are important in any market, especially financial markets, since they imply that interaction with other market participants is more important than simple cognitive ability the longer one participates in the market; anticipating the decisions and actions of other players is a function of experience, not necessarily cognitive ability. A possible extension to the present paper would be to see if individuals' trading behavior in financial markets varies based on their CRT results, which may confirm the previous conjecture. We would expect individuals with more extensive trading experience to more successfully incorporate other market participants' actions, whether their CRT scores are high or low. At this point, we can only speculate as to the relevance of our findings in the financial markets, but our results could be tested in that setting.

The question that remains to be answered is whether cognitive or impulsive users are better at incorporating other players' choices into their decision-making process. While System C users pick answers initially that are closer to the theoretical equilibrium of zero, this will not help them "win" the contest unless most of the other players are also System C users or unless they correctly predict, using reflection, how the other users in the contest play the game.

An interesting follow-up study would be to ask students directly whether they believe the other participants are impulsive or reflective. That is, how much noise do the participants expect to encounter when they play the game? Perhaps some evidence that System C users can play the game fairly well in round 1 comes from the observation that two of the three winners in the first round are classified as System C. In round 2, however, only two out of seven winners are classified as System C. The other winners are the "lucky" noise traders.

References

Burnham, T.C., D. Cesarini, M. Johannesson, P. Lichtenstein, and B. Wallace. 2009. "Higher Cognitive Ability is Associated with Lower Entries in a p-Beauty Contest. *Journal of Economics Behavior & Organization* 72(1): 171-175.

Camerer, C.F. and T.H. Ho. 2000. *Strategic Learning and Teaching*, California Institute of Technology Social Science Working Paper 1100.

Frederick, S. 2005. "Cognitive Reflection and Decision Making," *Journal of Economic Perspectives* 19(4): 24-42.

Keynes, J. M. 1936. *The General Theory of Employment, Interest and Money*, Harcourt Brace and Co.

Klein, G. 1999. *Sources of Power: How People Make Decisions*, MIT Press.

Montier, J. 2010. *The Little Book of Behavioral Investing*, John Wiley & Sons, Ltd.

---

[12] Of course, this is only a small experiment in one classroom setting. The results may differ substantially if the experiment is extended to the financial markets or if cognitive reflection is measured differently. Moreover, it is possible that System C users do not adjust their responses because they start with a better answer. Therefore, they have less reason to adjust their answer.

Moulin, H. 1986. *Game Theory for the Social Sciences,* 2nd Ed., NYU Press.

Nagel, R. 1995. "Unraveling in Guessing Games: An Experimental Study." *American Economic Review* 85(5): 1313-1326.

Obrecht, N.A., G.B. Chapman, and R. Gelman, 2007. "Intuitive t Tests: Lay Use of Statistical Information. *Psychonomic Bulletin & Review* 14(6): 1147-1152.

Pinillos, N.Á., N. Smith, G.S. Nair, P. Marchetto, and C. Mun. 2011. "Philosophy's New Challenge: Experiments and Intentional Action." *Mind and Language* 26(1): 115-139.

Thaler, R. 1998. "Giving Markets a Human Dimension," in *The Complete Finance Companion*, Financial Times/Prentice Hall.

Appendix

1. A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?

2. If it takes five machines five minutes to make five widgets, how long would it take 100 machines to make 100 widgets?

3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half the lake?

4. Pick a number from the range 0 to 100. The winner will be the person who picks the number closest to two-thirds of the average number picked.