Cognitive Ability and Behavior in the Lab: Are Differences in Intrinsic Motivation a Problem?

Matthew P. Taylor[†]

September 14, 2018

Abstract

I use an experiment to test whether economics experiments that have explored the relationship between cognitive ability and several important economic behaviors have biased estimates because they fail to account for the impact of differences in intrinsic motivation. I find that monetary incentives do not significantly improve subject performance on the types of questions that are commonly-used to measure cognitive ability. I also find that estimates of the relationships between cognitive ability and strategic reasoning, trust, and risk aversion are not significantly different whether cognitive ability is measured with or without monetary incentives. Consistent with the existing literature, subjects with higher cognitive ability demonstrate higher levels of strategic reasoning and they tend to be more trusting. However, in contrast to some prior studies, they are not more risk tolerant.

Keywords: decision making, cognitive ability, intrinsic motivation, risk, trust, reasoning

JEL classification: C91, D81, G41

[†]University of Montana, Department of Economics, email: matthew.taylor@mso.umt.edu).

1 Introduction

Economists have generated a substantial literature exploring the relationships between cognitive ability and decision making, such as investing behavior, time preferences, risk preferences, trust, and level of reasoning.[1] In the experimental studies that have explored these relationships, the typical protocol measures the economic behavior of interest with decisions that are incentivized with monetary payoffs, but does not incentivize performance on the tests used to measure cognitive ability.[2]

Although it does reduce the cost of an experiment, it is somewhat surprising that the standard protocol that economists have adopted to explore these relationships uses an unincentivized test to measure cognitive ability for two reasons. First, economists generally agree that incentives matter when a task requires effort (Camerer and Hogarth, 1999), and the tests that are used to measure cognitive ability certainly do. In fact, the cognitive reflection test (CRT), one of the most commonly used tests to measure cognitive ability, is specifically designed to measure cognitive processes that require "effort, motivation, concentration, and the execution of learned rules"(Frederick (2005), p.26). Second, there is empirical evidence that suggests that the effect of incentives on performance on ability tests is heterogenous and that incentives can increase the scores of individuals with low-baseline cognitive ability scores (i.e., scores on unincentivized tests) significantly more than individuals with higher baseline scores (Borgans et al., 2008; Heckman and Kautz, 2012).

Ability is a necessary condition to do well on these tests, but it is not sufficient. To do well, an individual must be motivated to solve these problems, and the standard protocol assumes subjects possess equal intrinsic motivation. However, when subjects of equal ability differ significantly in terms of their motivation to solve these puzzles their "ability" scores may differ substantially. The effects of these differences in intrinsic motivation can be significant. Duckworth et al. (2011) find that introducing incentives increases IQ scores nearly a full standard deviation for individuals with below-average baseline-IQ scores and nearly two-thirds of a standard deviation for the entire study sample.

The potential effects of ignoring the role of intrinsic motivation in our measures

[1] Branas-Graza and Smith (2016) catalogs this literature.

[2] A non-exhaustive list of studies that explore decision making and ability using non-incentivized tests include: Frederick (2005); Burnham et al. (2009); Campitelli and Labollita (2010); Cokely and Kelley (2009); Oechssler et al. (2009); Dohmen et al. (2010); Branas-Garza et al. (2012); Taylor (2013); Andersson et al. (2016); Corgnet et al. (2016); Cueva et al. (2016); Taylor (2016)

of ability are two-fold. First, it is possible that the estimated relationships between decision making and ability are actually being primarily driven by differences in intrinsic motivation. Individuals with high levels of intrinsic motivation may actually *enjoy* solving the problems, puzzles, and "brain-teasers" presented to them during the cognitive ability tests and decision-making tasks. Second, it is possible that the estimated relationships between preferences and ability are stronger and our estimates are biased downward because high-ability individuals with low intrinsic motivation are misidentified as low-ability individuals. This possibility, however assumes that monetary incentives improve performance and that may not necessarily be the case. Ariely et al. (2009) finds that large stakes for tasks that require creativity and problem solving can actually worsen performance, whereas Gneezy and Rustichini (2000) find that a small monetary payment worsened performance relative to no payment, but larger payments improved it—implying that the effect of incentives may be task dependent and, thus, they need to be calibrated to the task.

Recognizing that measuring ability with a unincentivized test could be biased, Branas-Garza et al. (2016) conducted a meta-analysis of 118 studies to explore whether CRT scores were higher in the fourteen percent of the studies that used incentivized CRT tests. Their results were inconclusive, and the authors highlighted that they lacked details on the magnitude of the incentives and the procedures used in these studies.

This study addresses this gap using an experiment that measures cognitive ability for every subject under incentivized and unincentivized conditions with two nine-item tests, each including a subset of six CRT questions, and then compares whether incentives increase performance on the ability test and whether the estimated relationship between ability and economic behavior depends on the incentive conditions under which ability was measured. I focus on three types of economic behaviors that are important to decision making and have been shown to be correlated to cognitive ability in prior studies: the level of reasoning or strategic sophistication, trust, and risk preferences.

Keynes (1936) famously articulated the critical role of strategic sophistication when he described how the decision making of stock traders was similar to the "beauty contest" games in the newspapers of the time. In fact, for most of us, not a day passes in which we do not make a decision that does not require us to consider the reasoning and strategy of another person. Experiments exploring the relationship between

cognitive ability and strategic sophistication have consistently found that individuals with higher cognitive ability demonstrate higher levels of reasoning. Branas-Garza et al. (2012), Carpenter et al. (2013), and Fehr and Huck (2016) find that subjects who perform better on Frederick's three-item CRT demonstrate higher levels of reasoning when playing a one-round p-beauty contest game (BCG). Burnham et al. (2009) also find similar results measuring ability with a 20-minute proprietary test of analogies, number series, and logical series. Gill and Prowse (2016) find that subjects who performed better on a 60-item nonverbal Raven Progressive Matrices played numbers closer to equilibrium and were more likely to converge to the Nash equilibrium during a 10-round BCG. All five studies incentivized the BCG, but only Fehr and Huck (2016) incentivized the test used to measure ability during the experiment (subjects earned 1 Euro for each correct response).

Most of our economic interactions involve trust as well, and recent studies have shown that individuals with higher levels of trust are more likely to participate in labor markets (Tu and Bulte, 2010), become an enterpreneur (Guiso et al., 2006), and own stocks and invest in risky financial assets (Guiso et al., 2008; Delis and Mylonidis, 2015). Two studies find that cognitive ability is positively correlated with trust. Using data from the General Social Survey, Carl and Billari (2014) find that performance on a 10-item vocabulary test is positively correlated with a commonly-used self-reported measure of generalized trust, neither measure was incentivized. Corgnet et al. (2016) find a positive relationship between unincentivized CRT scores and trust using an incentivized trust game.

Finally, many of the most important decisions we face, such as buying a home, choosing an occupation, deciding to change jobs, or deciding when to retire, involve risk. Individuals who self-report a greater willingness to take risks are more likely to be self-employed, be active in sports, and invest in stocks (Dohmen et al., 2011), and there is an emerging consensus that cognitive ability is inversely related to risk aversion (see Dohmen et al. (2018) for a recent summary). However, several experiments do not find a statistically significant relationship (Tymula et al., 2012; Taylor, 2016), another demonstrates that the estimate relationship can be reversed with an alternative risk preference task (Andersson et al., 2016), and one finds that it is only present for risk preferences elicited with a hypothetical task (Taylor, 2013).

I find that subjects do not perform significantly better on the cognitive ability test that is incentivized with monetary payments relative to the one that is not. Subjects

who perform relatively better when the test is incentivized also tend to perform relatively better when it is not. This lack of difference in performance holds for the subset of CRT questions, as well. In fact, the proportion of correct responses was greater for seven of the twelve CRT questions under unincentivized conditions than incentivized conditions, although not at a statistically significant level. Additionally, paying subjects did not significantly reduce the proportion of subjects who gave the impulsive response for the CRT questions.

I also find that estimates of the relationship between cognitive ability and strategic reasoning, trust, reciprocity, and risk aversion are not significantly different when cognitive ability is measured with an unincentivized test. Regardless of the conditions under which ability was measured, the results from this experiments support the following conclusions: (a) subjects with higher cognitive ability tend to average lower guesses in the beauty contest game and earn more money, (b) higher ability subjects tend to trust more than lower ability subjects when playing the trust game but that trust does not result in greater earnings, (c) cognitive ability is not correlated with reciprocity, and (d) cognitive ability is not correlated with risk preferences.

This paper proceeds as follows: the experimental design and description of the tasks is described in the next section. Empirical results are presented and discussed in Section 3, and Section 4 concludes.

2 Experimental Design

2.1 Method

One hundred and sixty subjects were recruited via email at the University of Montana from more than thirty different courses in more than fifteen different disciples that included astronomy, biology, chemistry, mathematics, philosophy, political science, and economics. The experiment was conducted in a computer lab on campus and was computer-based. There were twenty-four sessions between October 2017 and April 2018 and the average number of subjects in each session was 6.67, although every session had either four, eight, or twelve subjects, and subjects were separated by dividers. Subjects earned an average of $32.98 and each session lasted about one hour.

2.2 Procedure

Cognitive ability. All subjects completed two nine-item cognitive ability tests: Version A and Version B. Both versions included six CRT questions, two simple questions testing a subject's comprehension of percentages, and one conditional probability (CP) question. The order of the questions was randomized for each subject, except the CP question was always presented last. The text of each question in Version A is shown in the first column of Table 1. If the question is a CRT question the correct response and impulsive response are shown in Columns 2 and 3, respectively. Each question was given a short name, which is shown in Column 4, and the source of each question is in Column 5. The corresponding information for Version B is shown in Table 2.

Version A included the three original cognitive reflection test questions introduced by Frederick (2005) plus three additional CRT questions. Q4: HOLE and Q5: RACE were developed by Thomson and Oppenheimer (2016) and Q8: ELVES was introduced in Primi et al. (2016), but attributed to Shane Frederick. Q9: MAMMOGRAM was accompanied by a cross-tabulation table which is shown in Figure 10 in the Appendix.

Table 1: Ability Test—Version A Description

Question	Correct Response	Impulsive Response	Short Name	Source
Q1: 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?	5 cents	10 cents	Q1: BAT & BALL	Frederick (2005)[a]
Q2: In a lake, there is a patch of lilypads. 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?	47	24	Q2: LILYPAD	Frederick (2005)
Q3: If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?	5 minutes	100 minutes	Q3: MACHINES	Frederick (2005)[a]
Q4: How many cubic feet of dirt are there in a hole that is 3 feet deep x 3 feet wide x 3 feet long?	0	27	Q4: HOLE	Thomson and Oppenheimer (2016)
Q5: If you are running a race and you pass the person in second place, what place are you in?	second	first	Q5: RACE	Thomson and Oppenheimer (2016)
Q6: In the BIG BUCKS LOTTERY, the chances of winning a $10.00 prize are 1%. What is your best guess about how many people would win a $10.00 prize if 1,000 people each buy a single ticket from BIG BUCKS?	10	—	Q6: BIG BUCKS	Schwartz et al. (1997)[a]
Q7: If the chance of your flight departing late is 15 out of 100, this would be the same as having a __ % chance of a late departure.	15	—	Q7: FLIGHT	Author
Q8: If 3 elves can wrap 3 toys in an hour, how many elves are needed to wrap 6 toys in 2 hours?	3	6	Q8: ELVES	Primi et al. (2016)[b]
Q9: Suppose you have a close friend who has a lump in her breast and must have a mammogram. Of 100 women like her, 10 of them actually have a malignant tumor and 90 of them do not. Of the the 10 women who actually have a tumor, the mammogram indicates correctly that 9 of them have a tumor and indicates incorrectly that 1 of them does not have a tumor. Of the 90 who women who do not have a tumor, the mammogram indicates correctly that 81 of them do not have a tumor and indicates incorrectly that 9 of them do have a tumor. The table summarizes all of this information. Imagine that your friend tests positive (as if she had a tumor), what is the likelihood that she actually has a tumor? [Table shown to subjects in Appendix]	9/18	—	Q9: MAMMOGRAM	Peters et al. (2007)[a]

[a] Question included in eight-item cognitive ability test developed by Weller et al. (2013).
[b] Authors credit Shane Frederick as original source of question.

Version B included three CRT questions introduced in Baron et al. (2015), along with one from Thomson and Oppenheimer (2016), and two attributed to Shane Frederick and used by Primi et al. (2016). Q18: COLORBLIND was also accompanied by a cross-tabulation table, which is shown in Figure 11 in the Appendix.

All subjects completed both versions of the cognitive ability test. One of the tests was incentivized and they earned \$1 for each correct response. The other test was unincentivized. The version of the test that each subject did first and whether they were paid for the first or second test they completed was randomized and counterbalanced. Subjects were not informed that they would be completing a second test when they completed the first test. Although there are eight conditions under which a test could be completed because of the 3×2 factorial design, there are actually only four treatments since a subject could not complete the same version twice. For example, if a subject was randomly assigned to complete Version A under paid conditions for her first test, then she necessarily completed Version B under unpaid conditions for her second test. Subjects were not notified of the number of correct responses until the end of the experiment. Between each test subjects completed a choice task designed to measure a subject's risk and skewness preferences for which they did not learn the results until the end of the experiment. The results of the skewness task are not discussed in this paper.

Table 2: Ability Test—Version B Description

Question	Correct Response	Impulsive Response	Short Name	Source
Q10: If it takes 2 nurses 2 minutes to measure the blood pressure of 2 patients, how long would it take 200 nurses to measure the blood pressure of 200 patients?	2	200	Q10: NURSES	Baron et al. (2015)
Q11: Soup and salad cost $5.50 in total. The soup costs a dollar more than the salad. How much does the salad cost?	2.25	2.50	Q11: SOUP&SALAD	Baron et al. (2015)
Q12: Sally is making sun tea. Every hour, the concentration of the tea doubles. If it takes 6 hours for the tea to be ready, how long would it take for the tea to reach half of the final concentration?	5	3	Q12: TEA	Baron et al. (2015)
Q13: Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are there in the class?	29	30	Q13: CLASS	Primi et al. (2016)[b]
Q14: Emily's father has three daughters. The first two are named April and May. What is the third daughterś name?	Emily	June	Q14: EMILY	Thomson and Oppenheimer (2016)
Q15: Imagine that we roll a fair, six-sided die 1,000 times (That would mean that we roll one die from a pair of dice.). Out of 1,000 rolls, how many times do you think the die would come up as an even number?	500	—	Q15: DICE	Schwartz et al. (1997)[a]
Q16: If the chance of getting a disease is 20 out of 100, this would be the same as having a __ % chance of getting the disease.	20	—	Q16: DISEASE	Schwartz et al. (1997)[a]
Q17: On an athletics team, tall members are three times more likely to win a medal than short members. This year the team has won 60 medals so far. How many of these have been won by short athletes?	15	20	Q17: TEAM	Primi et al. (2016)[b]
Q18: Suppose that 5 percent of men and 0.25 percent of women are colorblind. A colorblind person is chosen at random out of a sample of 800 people. That is the likelihood of this colorblind person being male? [Table shown to subjects in Appendix]	20/21	—	Q18: COLORBLIND	Adapted from Ross (2002)

[a] Question included in eight-item cognitive ability test developed by Weller et al. (2013).
[b] Authors credit Shane Frederick as original source of question.

Reasoning. Subject's reasoning abilities were measured using a five-round, four-person p-beauty contest game (BCG) without rematching. In the BCG, which was operationalized by Nagel (1995), subjects attempt to choose the number on an interval from 0 to 100 that is some fraction, $p < 1$, of the all the numbers that the N number of subjects in the game have chosen. The Nash equilibrium of the game is for all the subjects to choose zero. However, most subjects do not choose zero and the value of p, which in this case equaled 0.5, can be used to identify the level k reasoning of a subject. For instance, a subject who choses 25 in the first round is said to be using level 1 reasoning, and another who chooses 12.5 is using level 2 reasoning.

The winner of each BCG round received $2.00 and ties were split evenly among the winners. Subjects had sixty seconds to submit their choices in each round and a countdown timer was displayed at the top of the page. Subjects were informed that if they failed to submit a number within sixty seconds they would receive $0 for the round. At the end of each round, subjects were provided with the following information: (i) the numbers chosen by all group members, (ii) the average of all four numbers, (iii) 50 percent of the average, (iv) which group member(s) won the round, and (iv) what the subject won for that round. Before the first round, subjects completed a set of instructions that explained the rules of the game, informed them of the number of rounds, what information they would be provided at the end of each round, and they were required to correctly answer a question that tested whether they understood how the winner of each round was determined.

Trust and reciprocity. Trust and reciprocity preferences were measured using the conventional trust game (Berg et al., 1995). The trust game is a two-player game that provides a measure of one player's willingness to trust and the other player's reciprocity. Individuals were matched within each session in the order that they completed the instructions for the task, but random otherwise. The first mover was endowed with $5 and chose how much to the cent, if any, to pass to the second mover. The amount that the first mover passed was tripled and the second mover decided how much to the cent, if any, to pass back. Each player had two minutes to make a decision.

Risk aversion. Risk aversion was measured with the conventional HL MPL shown in Table 3 (Holt and Laury, 2002). The HL MPL presents subjects with ten decisions between a "safe" lottery and a "risky" lottery and subjects indicate which lottery they prefer to play for each decision. The safe lottery in this experiment had potential

payoffs of \$6.00 and \$4.80 and the risky lottery had potential payoffs of \$11.55 and \$0.30 (3 × HL "baseline" payoffs). Half of the subjects saw the safe lottery on the left and half saw the safe lottery on the right, and the lotteries were referred to as Option A and Option B throughout the experiment. The probability of realizing the high and low payoffs in each decision are identical across the lotteries in the HL MPL, thus, for the first four decisions the expected value of the safe option is greater than the risky option. As the probability of the high payoff is increased in both options by ten percentage points the expected value of the risky payoff increases at a greater rate than the safe option and becomes greater at Decision 5. Hence, we can infer an individual's risk tolerance based on the number of safe choices she makes—fewer safe choices imply more risk tolerance and a risk-neutral subject will make four safe choices.

Table 3: Holt & Laury Multiple Price List

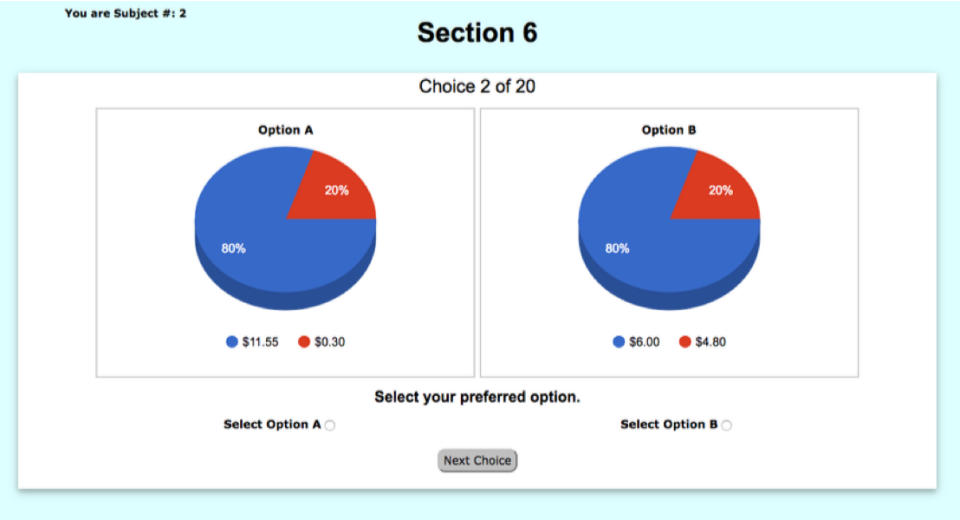
HL	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Decision	Option A		Option B					
number	p(\$6)	p(\$4.80)	p(\$11.55)	p(\$0.30)	EV^A	EV^B	$EV^A - EV^B$	CRRA interval
1	0.1	0.9	0.1	0.9	4.92	1.43	3.49	$r < -1.71$
2	0.2	0.8	0.2	0.9	5.04	2.55	2.49	$-1.71 < r < -0.95$
3	0.3	0.7	0.3	0.7	5.16	3.68	1.49	$-0.95 < r < -0.49$
4	0.4	0.6	0.4	0.6	5.28	4.80	0.49	$-0.49 < r < -0.15$
5	0.5	0.5	0.5	0.5	5.40	5.93	-0.53	$-0.15 < r < 0.15$
6	0.6	0.4	0.6	0.4	5.52	7.05	-1.53	$0.15 < r < 0.41$
7	0.7	0.3	0.7	0.3	5.64	8.18	-2.54	$0.41 < r < 0.68$
8	0.8	0.2	0.8	0.2	5.76	9.30	-3.54	$0.68 < r < 0.97$
9	0.9	0.1	0.9	0.1	5.88	10.43	-4.55	$0.97 < r < 1.37$
10	1	0	1	0	6.00	11.55	-5.55	$1.37 < r$

Notes: All currency units are in 2018 U.S. dollars. Subjects were not presented with the information in columns (5) through (8). Column (8) assumes utility function with constant relative risk aversion (CRRA), $u(w) = \frac{w^{(1-r)}}{(1-r)}$

Although all ten decisions are often presented simultaneously and in the order shown in Table 3, each decision was presented separately in this experiment, as shown in Figure 1, and the order was randomized. This was done because these decisions were combined with ten additional decisions derived from Drichoutis and Lusk (2016) for the first seventy-six subjects. The ten-decision DL MPL changes the higher payoffs across the decisions while holding the probabilities constant, but a subject's risk preference implied by the switch point is identical to the HL MPL (if the assumption of CRRA holds). However, only one of the seventy-six subjects who completed these twenty choices had an identical switch point across the two MPLs. Also, like Drichoutis and Lusk (2016), I find an extraordinary level of inconsistency

in subjects' choices on the DL MPL—more than 70 percent of these subjects made at least one inconsistent choice on these ten decisions, which makes it difficult to use the switch point from the DL MPL as a summary measure of a subject's risk preference. Interestingly, as I show in the Results section below, including the DL MPL decisions did not significantly affect the consistency of choices on the ten HL MPL decision, but it did affect the switch point.

Figure 1: HL MPL Screenshot



3 Results

3.1 Paid vs. Unpaid Measures of Cognitive Ability

On average, subjects did not perform significantly better on the ability tests under paid conditions, nor did they devote more time to answering them. The percentage of subjects who answered each question correctly under paid and unpaid conditions and the p-values of the one-sided test of the null hypothesis that the paid proportion correct is greater than the unpaid proportion correct are shown in Table 4. Subjects did not perform significantly better on a single question in the 18-item battery. In fact, subjects performed better under unpaid conditions on six of the nine questions in Version A and three out of nine of the questions in Version B, but not at a statistically

significant level.[3]

A comparison of total mean scores for Version A at the bottom of Panel A of Table 4 shows that subjects performed worse under paid conditions. Subjects average 4.48 correct out of nine possible under paid conditions and 4.88 correct under unpaid conditions. The pattern was similar for the subset of six CRT questions and the original three CRT questions—subjects average fewer correct responses on Version A under paid conditions. Panel B of Table 4 shows that subjects performed slightly better under paid relative to unpaid conditions, but not at a statistically significant level. They averaged 5.33 correct under paid conditions and 5.15 under unpaid conditions.

[3]For every question, two-sided tests of equal proportion fail to reject the null. For one-sided tests that subjects performed *better* under unpaid conditions, only Questions 1 and 8 have p-values of less than ten percent.

Table 4: Ability Test Summary Statistics

Panel A: Version A

Question	% Correct Paid	% Correct Unpaid	p-value paid >unpaid	% Impulsive Paid	% Impulsive Unpaid	p-value paid<unpaid
Q1: BAT & BALL	33.75	43.75	0.903	65.00	51.25	0.961
Q2: LILYPAD	46.25	52.50	0.785	38.75	38.75	0.500
Q3: MACHINES	27.50	35.00	0.847	57.50	46.25	0.923
Q4: DIRT	3.75	7.50	0.848	72.50	75.00	0.360
Q5: RACE	57.50	53.75	0.317	31.25	36.25	0.252
Q6: BIG BUCKS	86.25	81.25	0.196	—	—	—
Q7: FLIGHT	96.25	95.00	0.350	—	—	—
Q8: ELVES	62.50	72.50	0.912	30.00	23.75	0.814
Q9: MAMMOGRAM	30.00	37.50	0.842	—	—	—

	Mean Correct Paid	Mean Correct Unpaid	p-value paid>unpaid	p-value Wilcoxon Rank Sum	Mean Impulsive Paid	Mean Impulsive Unpaid	p-value paid<unpaid
Version A Score	4.48	4.88	0.907	0.237	—	—	—
All CRT questions	2.31	2.65	0.905	0.236	2.95	2.713	0.818
Frederick CRT score	1.08	1.31	0.914	0.189	1.613	1.363	0.938

Panel B: Version B

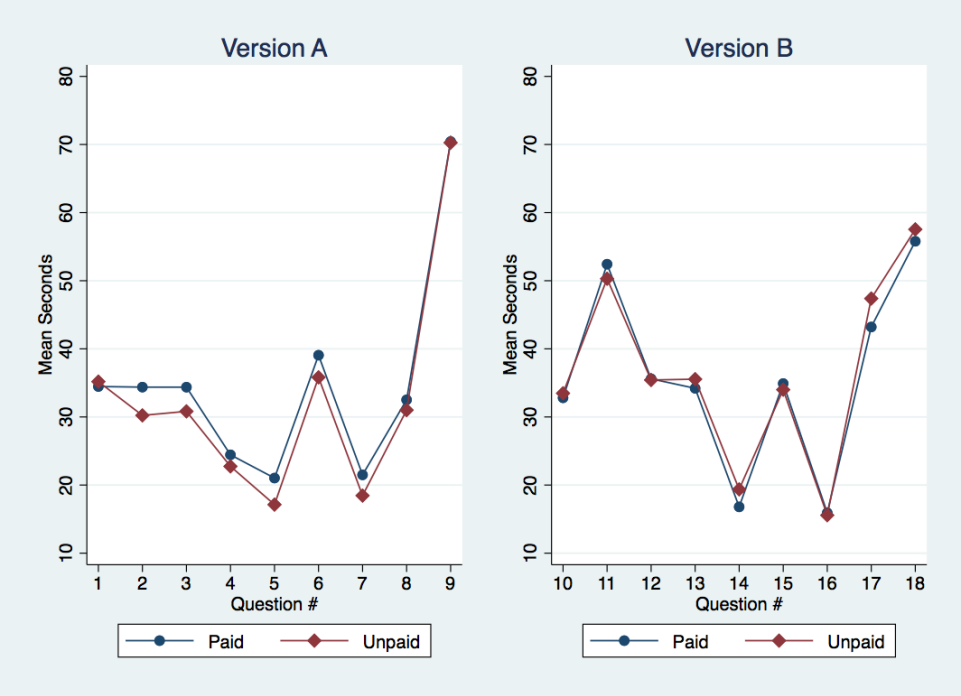
Question	% Correct Paid	% Correct Unpaid	p-value paid >unpaid	% Impulsive Paid	% Impulsive Unpaid	p-value paid<unpaid
Q10: NURSES	40.00	37.50	0.373	37.50	40.00	0.373
Q11: SOUP & SALAD	68.75	62.50	0.203	3.75	2.50	0.675
Q12: TEA	55.00	52.50	0.376	38.75	38.75	0.500
Q13: CLASS	23.75	26.25	0.643	42.50	50.00	0.171
Q14: EMILY	68.75	70.00	0.568	27.50	30.00	0.363
Q15: DICE	92.50	91.25	0.386	—	—	—
Q16: DISEASE	98.75	100	0.842	—	—	—
Q17: TEAM	46.25	41.25	0.262	47.50	43.75	0.683
Q18: COLORBLIND	40.00	33.75	0.206	—	—	—

	Mean Correct Paid	Mean Correct Unpaid	p-value paid>unpaid	p-value Wilcoxon Rank Sum	Mean Impulsive Paid	Mean Impulsive Unpaid	p-value paid<unpaid
Version B Score	5.34	5.15	0.285	0.545	—	—	—
All CRT questions	3.03	2.90	0.329	0.625	1.98	2.05	0.338

Notes: p-values shown for test of equality of proportions for individual questions.

Using time spent as a measure of effort, Figure 2 provides a comparison of time spent on paid and unpaid questions. This visual comparison strongly suggests that subjects did not devote significantly greater effort to these questions when they were paid. T-tests for each question of the null hypothesis that subjects spent more time on the paid questions support this conclusion. Only the test for Q5: RACE resulted in a p-value of less than 0.10.

Figure 2: Mean Time Spent on Each Question Under Paid and Unpaid Conditions

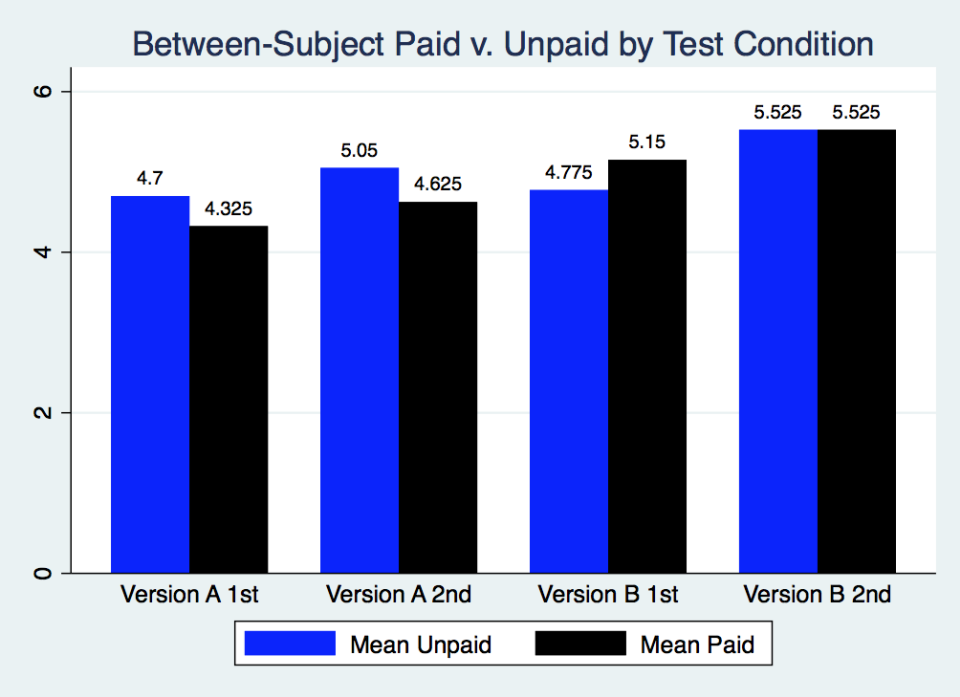


Incentives did not reduce the proportion that provided the impulsive responses for the CRT questions either. The proportion of impulsive responses was lower for five of the twelve CRT questions under paid and none of these proportions was significantly lower. At 0.171, Q13: CLASS has the lowest p-value for the test of the one-sided null hypothesis that the proportion providing the impulsive response under paid conditions was lower than under unpaid conditions.

Figure 3 provides a comparison of mean performance under the eight different testing conditions. When subjects completed Version A first and were not paid for the number of correct responses they answered 4.70 questions correctly, on average. When subjects completed Version A first and were paid the average *decreased* to 4.35. A similar pattern resulted when subjects completed Version A second—those who

were paid for the number of correct responses performed worse, on average, than those who were not paid. Subjects performance improved slightly when they were paid on Version B when it was completed first, but there was no difference in performance when it was completed second.

Figure 3: Number of Correct Responses—Paid versus Unpaid Ability Scores by Test Condition



To test how incentivizes and testing conditions affected subject performance, subjects' scores were modeled as a function of the triple interaction of dummy variables for *Version B*, the *Paid* condition, and the if the test was completed *Second* and estimated with random-effects panel regressions. Table 5 shows the results from these models for three different measures of the ability score: the 9-item score (*Score*); the 6-item CRT score (*CRT*), which includes the six CRT questions only; and, a 7-item score (*7-item*), which includes the six CRT questions and the conditional probability question, but excludes the two basic questions in each test.[4] The regressions with *CRT* and *7-item* as the dependent variable include only the subjects who answered the two basic questions correctly under the paid condition.[5]

[4]These are the two questions about simple percentages in each test—Q6: BIG BUCKS and Q7: FLIGHT in Version A, and Q15: DICE and Q16: DISEASE in Version B.

[5]All of three models were also estimated as panel Poisson regressions with all 320 observations

The coefficient estimates on *Paid* for all three regressions indicate that incentives did not improve performance on Version A when it was completed first, regardless of the ability measure used. This is unsurprising since Figure 1 shows that only when subjects completed Version B first was performance better when subjects were paid relative to unpaid (5.15 versus 4.775). However, even that improvement is not statistically significant. Wald tests of the appropriate combination of coefficients in Column (1) for each of the other three testing conditions are shown in the bottom panel of Table 5. All three tests fail to reject the null that these combined coefficients equal zero, which means that incentives did not improve performance under any set of testing conditions.

3.2 Within-Subject Comparisons of Performance under Paid and Unpaid Conditions

3.2.1 Order and Version Effects

Although the results above show that there are not average improvements in performance due to monetary incentivizes, it is possible that there are some individual subjects who may be extrinsically motivated. However, measuring ability under both paid and unpaid conditions necessarily required the use of two different test versions and may have resulted in order and version effects. Using the results from the panel regression shown in Column (1) of Table 5, Table 6 summarizes subjects' performance across the two versions and the order in which the test was completed and includes the p-values of Wald tests for the appropriate combination of coefficients from that regression. There is both an order effect and a version effect.

Subjects tended to perform better on the Second test they completed. For example, subjects who completed Version B under unpaid conditions as the first test averaged 4.775 correct responses, but subjects who completed Version B under unpaid conditions as the second test averaged 5.525 correct responses. Although the p-values comparing whether these order effects for each testing condition separately are not statistically different from zero, a parsimonious model that includes only the main effects of each condition indicates that subjects answered 0.44 more questions correctly on the test they completed second (p-value=0.001).[6]

and the results are qualitatively the same and are shown in Table 14 in the Appendix.

[6]P-values of the Wald tests generated from the panel regression shown in Column (1) of Table 5.

Table 5: Ability Scores and Treatment Conditions

	(1) Score	(2) CRT	(3) 7-item
Version B	0.075	-0.272	-0.167
	(0.446)	(0.384)	(0.426)
Paid	-0.375	-0.360	-0.343
	(0.428)	(0.346)	(0.368)
Version B × Paid	0.750	0.551	0.672
	(0.606)	(0.532)	(0.570)
Second Test	0.350	-0.378	0.027
	(0.468)	(0.409)	(0.431)
Version B × Second Test	0.400	1.320**	1.061*
	(0.654)	(0.560)	(0.617)
Paid × Second Test	-0.050	0.172	-0.027
	(0.606)	(0.534)	(0.557)
Version B × Paid × Second Test	-0.325	-0.519	-0.467
	(0.447)	(0.417)	(0.437)
Constant	4.700***	2.919***	3.108***
	(0.329)	(0.270)	(0.295)
Observations	320	284	284
r2_o	0.040	0.060	0.060
r2_w	0.217	0.196	0.248
r2_b	0.008	0.033	0.021

Wald Test: Paid vs. Unpaid	p-value
Version A, Second	0.321
Paid + Paid×Second Test=0	
Version B, First	0.382
Paid + Paid × Version B=0	
Version B, Second	0.999
Paid + Paid×Second + Paid×Version B + Version B×Paid×Second=0	

Standard errors clustered on subject in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Comparison of Mean Scores Across Conditions

			Test Version		
			Test A		**Test B**
				← p-value →	
		First	4.700	*0.866*	4.775
	Unpaid	↑ p-value ↓	*0.454*		*0.101*
		Second	5.050	*0.321*	5.525
Incentives					
		First	4.325	*0.044*	5.150
	Paid	↑ p-value ↓	*0.436*		*0.339*
		Second	4.625	*0.046*	5.525

Notes: p-values estimated using a random-effects panel regression model with standard errors clustered on subject.

The version effect is more subtle. Subjects performed better on Version B than Version A under unpaid conditions, but not at a statistically significantly level. However, subjects performed better on Version B than Version A under paid conditions at a 0.05 level of statistical significance. This difference is due to worsened performance on Version A when subjects were paid, but improved performance on Version B when they were paid. That is, incentivizes did not cause performance to worsen or improve significantly on either version of the test, but they had divergent effects on performance, so average performance was significantly better on Version B than Version A under Paid conditions.

Given these order and version effects, it is necessary to standardized the ability scores to explore whether intrinsic motivation is a problem when considering the relationship between ability and preferences. Thus, the ability scores used below are standardized scores based on the mean and standard deviation in each of the eight testing conditions (3×2 factorial design). These standardized scores, which measure how subjects performed relative to other subjects under the same testing conditions, allow within-subject comparisons for each subject under Paid and unpaid conditions .

3.2.2 Identifying Extrinsically-Motivated Subjects

To determine whether unpaid ability scores are biased because of differences in motivation, it will be useful at times in the next section to use the standardized scores to categorize subjects based on their performance on both the paid and unpaid tests. The scatter plot of subjects' *Std. Paid Score* and *Std. Unpaid Score* in Figure 4 suggests that performance under paid and unpaid conditions is strongly correlated, and it is—the correlation coefficient for the two scores is 0.755. Subjects with standardized scores above zero in both the paid and unpaid tests are defined as *High Ability (HA)*. Subjects with standardized scores below zero in both conditions are defined as *Low Ability (HA)*. Subjects who score below zero on the unpaid test but greater than zero on the paid test are defined as *Externally-Motivated (EM)*, and subjects who score above zero on the unpaid test but below zero on the paid test are defined as *Stressed*.

Sixty-five (40.6%) subjects can be categorized as *HA* (Quadrant I), seventeen (10.5%) can be categorized as *EM* (Quadrant II), sixty-one (38.1%) as *LA* (Quadrant III), and seventeen (10.6%) as *Stressed* (Quadrant IV). Nearly 80% of the sample scored either above the mean in both paid and unpaid conditions or below the mean in both conditions.

While categorizing subjects in the above way is primarily useful for visual comparisons, an alternate measure of external motivation that will be used in the regression analysis in the next section is the difference between *Std. Paid Score* and *Std. Unpaid Score*, denoted *Std. Dif. Paid Unpaid*. A larger *Std. Dif. Paid-Unpaid* indicates a more externally-motivated subject.

3.3 Beauty Contest Game

Whereas an individual's willingness to trust, reciprocate, or take risks are preferences, her ability to consider the reasoning and behavior of other players is a cognitive skill. Thus, of the three characteristics considered in this study, a subject's behavior in the beauty contest game is the most likely to be correlated with subject performance on the cognitive tests. Figure 5 shows that *HA* subjects made lower guesses than subjects in the other groups in every round, *LA* subjects tended to make the highest guesses, and the average guesses of the subjects in the *EM* and *Stressed* categories fell between the averages of the other two groups, except the *Stressed* subjects made

Figure 4: Within-Subject Standardized Scores

the highest guesses in the first round.

Figure 6 shows that these lower guesses resulted in consistently higher earnings for *HA* subjects. Interestingly, in the first round, *EM* subjects earned slightly more, on average, than *HA* subjects. Otherwise, however, *EM* subject's per-round earning fall between the higher earning *HA* subjects and the lower earning *LA* subjects.

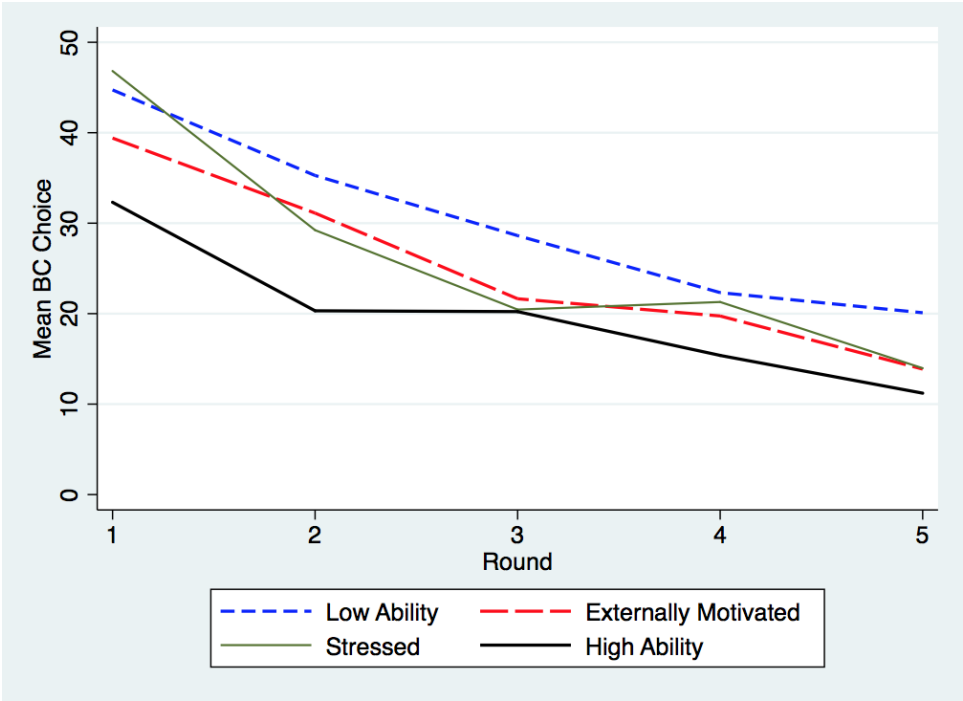
The fact that EM guesses and earnings tend to fall between the HA and LA subjects' guesses could be the result of this category being a mixture of high ability subjects and low-ability subjects, which implies that estimates of the relationship between ability and performance are downward biased if the unincentivized measure of ability is used. It could also simply be that EM subjects are of mediocre ability.

To test whether motivational differences are confounding our understanding of the relationship between ability and BCG performance, I estimate three different versions of the following equation for each of the five rounds separately:

$$BC_i = \beta_0 + \beta_1 * Std.\ Score + \beta_2 * Std.\ Dif.\ Paid - Unpaid + \Pi * X_i + \epsilon_i \quad (3.1)$$

where *BC* denotes the subject's guess for that round, *Std. Score* denotes either the Unpaid or Paid standardized score, *Std. Dif. Paid-Unpaid* captures external

Figure 5: Round-by-Round Mean Guesses by Ability Group

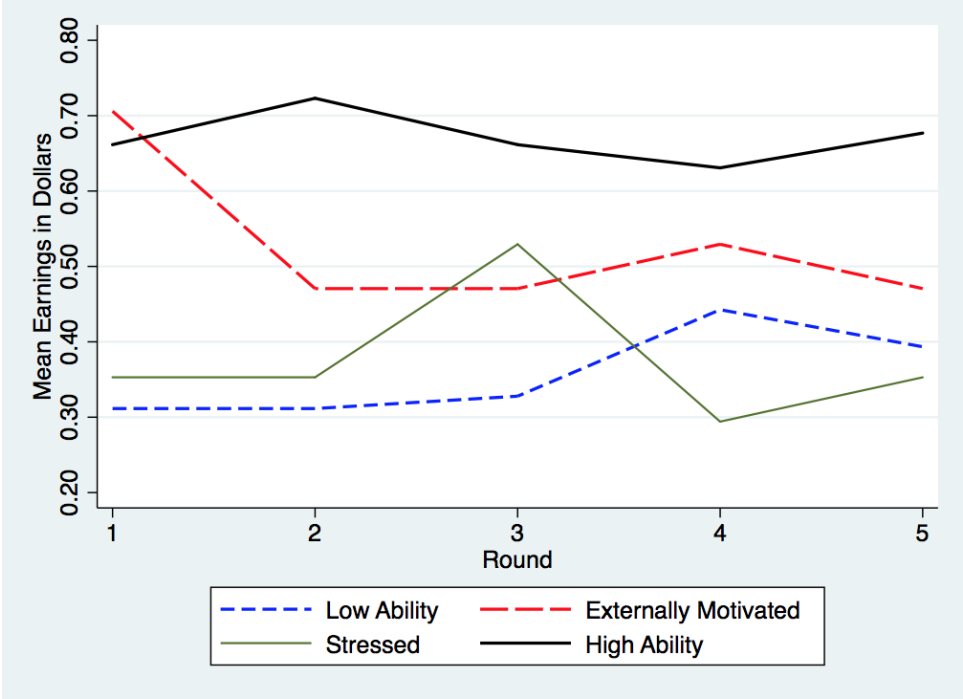


motivation and is the difference between *Std. Paid Score* and *Std. Unpaid Score*, X_i denotes a set of controls for whether the subject came to the experiment with a friend (*Friend in Session*), gender (*Female*), risk aversion (*HL MPL Switch Point*), if he or she was an economics major (*Economics Major*), and the number of attempts on the BCG Quiz (*BCG Quiz Attempts*).[7]

The resulting coefficient estimates on the relevant standardize score variables from those fifteen regressions are shown in Table 7. The first specification includes the *Std. Unpaid Score* as the *Std. Score* and is shown in Panel A. The second replaces the unpaid score with *Std. Paid Score* and is shown in Panel B. The third specification, shown in Panel C, includes *Std. Unpaid Score* and the measure of external motivation, *Std. Dif. Paid-Unpaid*. If intrinsic motivation is confounding the unpaid measure of ability, then the coefficient estimates on *Std. Unpaid Score* in Panel A of Table 7 will be significantly less in magnitude than those in Panel B on *Std. Paid Score*. We should also find the coefficient estimates on *Std. Dif. Paid-Unpaid* in Panel C to be negative and significant as well.

[7]This was a one question quiz that verified subjects understood how the winner of each round was determined.

Figure 6: Round-by-Round Mean Earnings by Ability Group



Panel A shows that in every round subjects with higher unpaid scores made significantly lower guesses, on average. If high-ability subjects appear to be low-ability based off unpaid scores because they are externally motivated, then the coefficients on *Std. Unpaid Score* will underestimate the "true" relationship between ability and depth of reasoning. Panel B shows the coefficient estimates when *Std. Unpaid Score* is replaced with *Std. Paid Score*. A comparison of the coefficient estimates in Panels A and B and Wald tests of the null that the coefficients are equal, shown in Panel D, provides evidence that they are not significantly different. The coefficient estimates on *Std. Unpaid Score* in Panel C resulting from the model that includes the measure of external motivation are similar. Additionally, the coefficient estimates on *Std. Dif. Paid-Unpaid* are consistently negative in all five rounds, which is what we would expect if high-ability subjects were being mistaken as low-ability subjects, but external motivation only has a statistically significant effect on guesses in the first round. Taken as a whole, although there may be some downward bias in the estimated relationship between ability and depth of reasoning if unpaid scores are used, these results indicate that it appears to be relatively minor.

Table 8 shows coefficient estimates from eighteen comparable regressions in which

Table 7: Beauty Contest Guesses and Ability Regression Results

	(1) Round 1	(2) Round 2	(3) Round 3	(4) Round 4	(5) Round 5
Panel A					
Std. Unpaid Score	-4.820***	-6.681***	-4.661***	-5.862***	-5.526***
	(1.859)	(1.509)	(1.534)	(1.355)	(1.245)
Panel B					
Std. Paid Score	-7.075***	-6.611***	-3.909**	-5.323***	-4.562***
	(1.888)	(1.415)	(1.658)	(1.757)	(1.574)
Panel C					
Std. Unpaid Score	-6.886***	-7.638***	-4.909***	-6.416***	-5.776***
	(2.016)	(1.554)	(1.676)	(1.631)	(1.506)
Std. Dif Paid-Unpaid	-7.639***	-3.535	-0.917	-2.049	-0.925
	(2.682)	(2.151)	(2.615)	(2.958)	(2.356)
Observations	158	158	158	158	158
Panel D					
Wald Tests: Std. Unpaid Score = Std. Paid Score					
Wald F-statistic	1.426	0.002	0.205	0.094	0.375
Wald p-value	0.234	0.960	0.651	0.759	0.541

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Two subjects dropped because of a non-response on either *Female* or *Friend in Session*.

the dependent variable is a subject's earnings for the indicated round and the total for all five rounds. Given that higher-ability subjects averaged lower guesses, it is unsurprising that these results show that they earn significantly more on average per round, but the relationship between standardized scores, regardless of the measure used, and earnings are only significant for the first three rounds. Moreover, there are not significant differences in the estimates that use *Std. Unpaid Score* and *Std. Paid Score*, except for Round 1, which shows that external motivation is associated with higher earnings in that round.

The cumulative effect of these per round differences in earnings resulted in rather large differences in total earnings for the BCG. Column 6 of Table 8 shows that a one standard deviation increase in the ability score resulted in approximately \$0.60 an additional earnings, regardless of the measure of ability used in the model. Given the scale of the payoffs, mean earnings are \$2.50 by design and the median is \$2.00, an increase of \$0.60 is practically significant. Indeed, the median earnings of subjects

Table 8: Beauty Contest Earnings and Ability Regression Results

	(1) Round 1	(2) Round 2	(3) Round 3	(4) Round 4	(5) Round 5	(6) All Rounds
Panel A						
Std. Unpaid Score	0.152**	0.138*	0.162**	0.063	0.046	0.560***
	(0.072)	(0.068)	(0.067)	(0.071)	(0.073)	(0.179)
Panel B						
Std. Paid Score	0.195***	0.142**	0.141**	0.103	0.012	0.594***
	(0.072)	(0.068)	(0.067)	(0.071)	(0.073)	(0.179)
Panel C						
Std. Unpaid Score	0.200**	0.161**	0.174**	0.096	0.032	0.664***
	(0.077)	(0.076)	(0.073)	(0.080)	(0.074)	(0.193)
Std. Dif. Paid-Unpaid	0.179*	0.086	0.044	0.124	-0.048	0.385
	(0.107)	(0.097)	(0.092)	(0.094)	(0.111)	(0.281)
Observations	158	158	158	158	158	158
Panel D						
Wald Tests: Std. Unpaid Score = Std. Paid Score						
Wald F-statistic	0.358	0.004	0.094	0.323	0.211	0.037
Wald p-value	0.551	0.949	0.760	0.570	0.646	0.849

Robust standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Two subjects dropped because of a non-response on either *Female* or *Friend in Session*.

categorized as *HA* were twice as large as the subjects in all three other categories ($4.00 v. $2.00).

3.4 Trust Game

Although the trust game was played within the confines of a single lab, the average amount passed and percentage of the available amount returned were consistent with prior experiments using the trust game (Berg et al., 1995; Cox, 2004; Garbarino and Slonim, 2009).[8] Player 1 subjects passed $3.00 to Player 2 subjects, on average. Player 2 subjects responded by returning approximately one-third of what they received, or $3.23, on average. This behavior resulted in Player 2 subjects earning about $0.66 more than Player 1 subjects—$5.21 versus $5.87.

Similar to the analysis of the BCG above, the relationship between trust and ability is modeled in three ways. Table 9 includes the results from these three regressions. All three models also include controls for whether the subject came to the experiment

[8]Cagno and Sciubba (2010) find that establishing social relationships with a network game prior to trust game play did not increase trust, which indicates having the subjects play in the same room should not significantly affect average play.

with a friend (*Friend in Session*), gender (*Female*), risk aversion (*HL MPL Switch Point*), if he or she was an economics major (*Economics Major*), and the number of attempts on the TG Quiz (*TG Quiz Attempts*).[9]

All three regressions indicate that a one standard deviation increase in the standardized ability score resulted in Player 1 subjects passing at least an additional $0.30, or nearly ten percent more than the average amount passed. Moreover, the coefficient estimates on *Std. Unpaid Score* in Column (1) in Table 9 and *Std. Paid Score* in Column (2) are nearly identical, as verified by the F-statistic and p-value from the Wald test of the null hypothesis that the coefficients are equal shown at the bottom of the column. Controlling for external motivation in the regression, as shown in Column (3), does not significantly change the coefficient estimate on *Std. Unpaid Score*. Higher ability players are more trusting and it does not matter whether ability is measured with an incentivized or unincentivized test.[10]

Also of note, the males in the sample are more trusting than female—females pass nearly $1.00 less than males on average. This is a common result (Buchan et al., 2008; Croson and Gneezy, 2009; Garbarino and Slonim, 2009; Dittrich, 2015), although some studies do not find a significant difference (it is rare to find females more trusting than males). Croson and Gneezy (2009) provide a summary of the literature and speculate that females may be more sensitive to differences in testing conditions than males, but there are no studies testing this explanation, or any other, to my knowledge.

To explore whether ability is related to reciprocity, the proportion that Player 2 returned to Player 1 is modeled as the dependent variable in the same three regressions with the amount passed by Player 1 (*Player 1 Passed*) included as an additional control. The results of these regressions, which are shown in Table 10, indicate that ability is unrelated to reciprocity, regardless of the measure used to control for ability. Corgnet et al. (2016) also did not find a relationship between ability and reciprocity.

The fact that coefficient estimates on *Player 1 Passed* in the Table 10 results were not significantly different from zero imply that the tendency for high-ability Player 1 subjects to pass greater amounts may have increased their earnings because Player 2 had a greater amount available. However, the results in Column (1) of Table 11, which model the TG earnings of Player 1, indicate that higher ability subjects did

[9]This was a one question quiz that verified subjects understood how the multiplier affected the amount Player 1 passed and how much Player 2 would have available to potentially pass back.

[10]These results are robust to using a Tobit model with an upper-limit.

Table 9: Trust Game: Player 1 Regression Results—Trust

	(1)	(2)	(3)
Std. Unpaid Score	0.327**	—	0.365**
	(0.161)		(0.163)
Std. Paid Score	—	0.322**	—
		(0.143)	
Std. Dif. Paid-Unpaid	—	—	0.144
			(0.200)
Friend in Session	0.053	0.062	0.054
	(0.357)	(0.362)	(0.356)
Female	-0.968***	-0.970***	-0.955***
	(0.321)	(0.328)	(0.326)
HL MPL Switch Point	-0.015	-0.038	-0.022
	(0.088)	(0.089)	(0.090)
Economics Major	0.092	0.023	0.027
	(0.451)	(0.454)	(0.463)
TG Quiz Attempts	-0.364	-0.267	-0.323
	(0.233)	(0.259)	(0.243)
Constant	3.536***	3.684***	3.570***
	(0.577)	(0.561)	(0.587)
Observations	79	79	79
Log-likelihood	-129.239	-129.508	-129.071
R-sq.	0.216	0.211	0.219
Wald Tests: Std. Unpaid Score = Std. Paid Score			
Wald F-statistic	—	0.001	0.056
Wald p-value	—	0.975	0.813

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

One subject dropped because of a non-response on *Friend in Session*.

Table 10: Trust Game: Player 2 Regression Results—Reciprocity

	(1)	(2)	(3)
Std. Unpaid Score	0.002	—	-0.002
	(0.027)		(0.027)
Std. Paid Score	—	-0.007	—
		(0.024)	
Std. Dif. Paid-Unpaid	—	—	-0.015
			(0.038)
Player 1 Passed (in dollars)	-0.004	-0.004	-0.004
	(0.024)	(0.024)	(0.024)
Friend in Session	0.092	0.090	0.092
	(0.096)	(0.094)	(0.097)
Female	-0.103*	-0.103*	-0.103*
	(0.056)	(0.057)	(0.057)
HL MPL Switch Point	0.003	0.004	0.004
	(0.009)	(0.009)	(0.009)
Economics Major	-0.050	-0.045	-0.047
	(0.075)	(0.073)	(0.075)
TG Quiz Attempts	-0.059**	-0.061**	-0.058**
	(0.025)	(0.027)	(0.025)
Constant	0.425***	0.421***	0.420***
	(0.129)	(0.130)	(0.130)
Observations	77	77	77
Log-likelihood	15.650	15.685	15.761
R-sq.	0.096	0.097	0.099
Wald Tests: Std. Unpaid Score = Std. Paid Score			
Wald F-statistic	—	0.131	0.024
Wald p-value	—	0.718	0.878

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Seventy-eight subjects were passed a positive amount from Player 1.

One subject dropped because of a non-response on *Female*.

not earn significantly more for themselves despite the tendency to trust more. Player 1 subjects earnings were also unaffected by the ability and external motivation of Player 2 subjects. In contrast, Player 2 subjects' earnings, modeled in Column (2), were greater if they played with higher ability subjects because high-ability Player 1 subjects tended to pass more. Player 2 subjects did not need to be smart, they just needed to play with a smart player.

Table 11: Trust Game: Regression Results—Earnings

	(1) Player 1 Earnings	(2) Player 2 Earnings
Std. Unpaid Score	0.121	0.202
	(0.247)	(0.488)
Std. Dif. Paid-Unpaid	0.411	0.277
	(0.379)	(0.563)
OP Std. Unpaid Score	0.036	0.730*
	(0.260)	(0.389)
OP Dif. Paid-Unpaid	0.283	0.269
	(0.403)	(0.525)
Friend in Session	-0.594	0.382
	(0.696)	(1.344)
Female	-0.257	-0.147
	(0.543)	(0.913)
HL MPL Switch Point	-0.198	-0.050
	(0.166)	(0.162)
Economics Major	-0.733	-1.056
	(1.200)	(0.877)
TG Quiz Attempts	-0.159	-0.142
	(0.399)	(0.582)
Constant	6.778***	6.298***
	(1.125)	(1.241)
Observations	78	79
Log-likelihood	-163.233	-204.205
R-sq.	0.066	0.062

Robust standard errors in parentheses

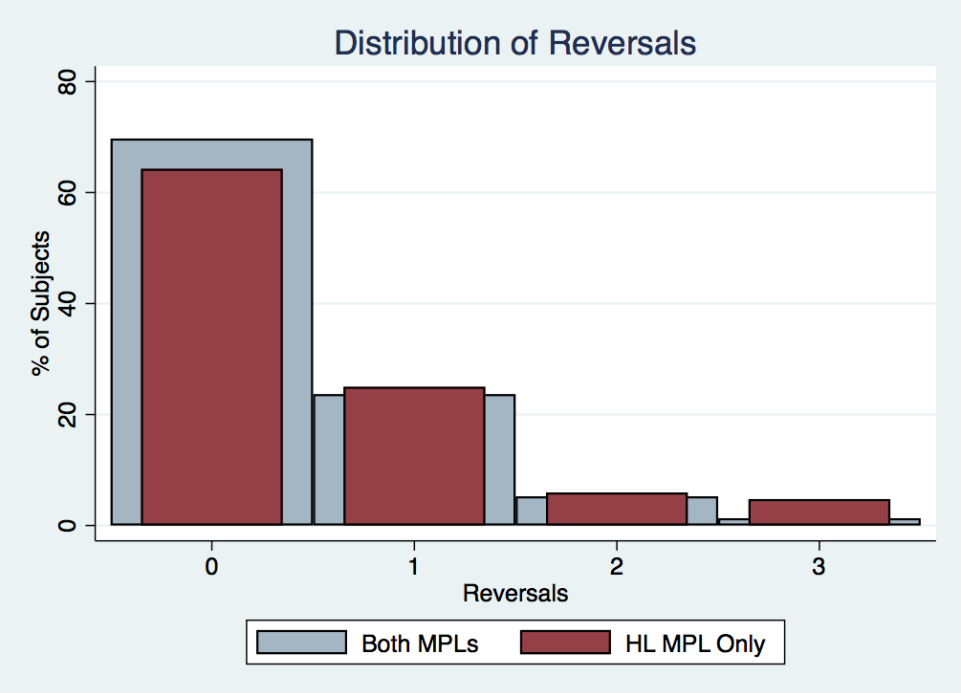
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.5 Risk Preferences

3.5.1 Inconsistency: Reversals

The randomization of the MPL appears to have affected the consistency of subjects' choices. Approximately one-third of the sample selected the safe option after having selected the risky option, whereas a sequential MPL typically only results in about ten to fifteen percent of the sample making *Reversals*.[11] Figure 7 depicts the distribution of reversals on the HL MPL and shows that combining the DL MPL and HL MPL did not cause subjects to make more reversals on the HL MPL. This is verified empirically in the logit regression results shown in Table 12, in which an indicator variable equal to one if a subject made at least one reversal, and zero otherwise, is regressed on the indicated standardized score, a dummy variable for female (*Female*), a dummy variable for economics majors (*Economics Major*), and a dummy for those subjects who completed the HL MPL only (*HL MPL Only*).[12].

Figure 7: Distribution of Reversal Frequency



The coefficient estimates on the standardized scores in all three regressions in

[11]For example, Holt and Laury (2002) find that thirteen percent of subjects commit at least one reversal in the low-payoff treatment and about six percent in the high-payoff treatment.

[12]Logit regressions results using CRT shown in Appendix Table 21

Table 12 are not statistically significantly different from zero, although they are consistently negative. This is somewhat surprising since Dave et al. (2010), Andersson et al. (2016), and Taylor (2016) all find that low-ability subjects are significantly more likely to make inconsistent choices, but those studies did not randomize the order of the HL MPL.

Table 12: Ability and Inconsistency

Logit regressions in which the dependent variable is an indicator variable equal to one if a subject made at least one reversal.			
	(1)	(2)	(3)
Std. Unpaid Score	-0.268	—	-0.291
	(0.175)		(0.187)
Std. Paid Score	—	-0.238	—
		(0.174)	
Std. Dif. Paid-Unpaid	—	—	-0.087
			(0.264)
Female	0.224	0.227	0.218
	(0.353)	(0.353)	(0.355)
Economics Major	-0.494	-0.490	-0.473
	(0.552)	(0.548)	(0.555)
HL MPL Only	0.231	0.222	0.221
	(0.346)	(0.344)	(0.343)
Constant	-0.892***	-0.888**	-0.887***
	(0.341)	(0.346)	(0.342)
Observations	159	159	159
Log-likelihood	-98.649	-98.896	-98.594
Wald Tests: Std. Unpaid Score = Std. Paid Score			
Wald Chi-sq stat.	—	0.030	0.016
Wald p-value	—	0.862	0.899

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

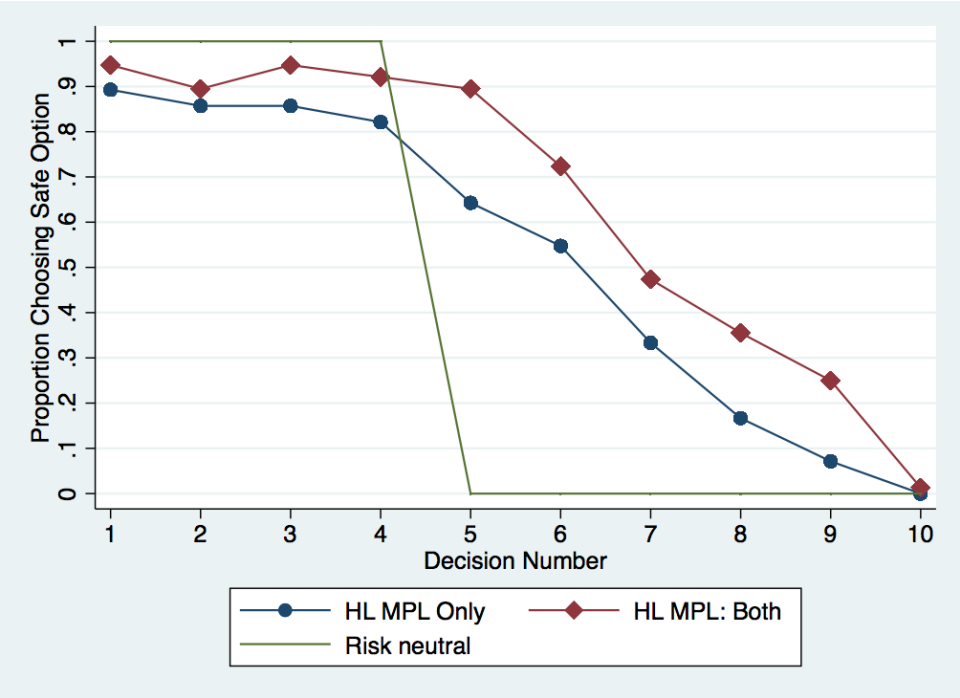
One subject dropped because of a non-response on *Female*.

3.5.2 Risk Aversion: Safe Choices & Switch Points

This sample, like most samples who complete the HL MPL, is risk averse. Figure 8 shows the proportion of subjects who chose the safe choice for each Decision under

the *HL MPL Only* treatment and the *Both MPLs* treatment. It also shows how risk-neutral subjects would choose. Interestingly, although mixing the DL MPL with the HL MPL did not cause in an increase in the number of subjects who made inconsistent choices, it is clear from a visual comparison of the two plots that a greater proportion of subjects chose the safe option for each HL MPL choice when the DL MPL choices were mixed in. In fact, twenty-five percent of the subjects in the *Both MPLs* treatment chose the safe option for Decision 9, which is an unusually high level of risk aversion on the HL MPL with this scale of payoffs.

Figure 8: Proportion of Safe Choices for Each Decision



One of the reasons the HL MPL is so popular in the experimental literature is that the subject's *Last Safe Choice* (i.e, the switch point) is a simple summary measure of each subject's risk aversion.[13] The mean *Last Safe Choice* of the sample is 6.35, and Figure 9 shows the distribution of subjects' *Last Safe Choice* in each treatment. Like Figure 8, a comparison of switch point distributions highlights the significantly

[13]Poisson model results using the nine-item score and the CRT score shown in Appendix Tables 22 and 23, respectively. Maximum likelihood estimation can be used to estimate a variety of different parameters related to decision making under risk or uncertainty with the primary limitations being the choices in the experiment and the sample size (Harrison and Rutström, 2008). However, Harrison (2006) demonstrated that parameter estimates can vary significantly depending upon the assumptions of the underlying model.

greater percentage of subjects who selected the safe option for Decision 9 when the MPLs were combined. A t-test of the null hypothesis that the mean switch point of the two treatments is equal is rejected at the one-percent level (p-value = .001), so a control for treatment is included in the regressions below.

Figure 9: Distribution of Subject Switch Points

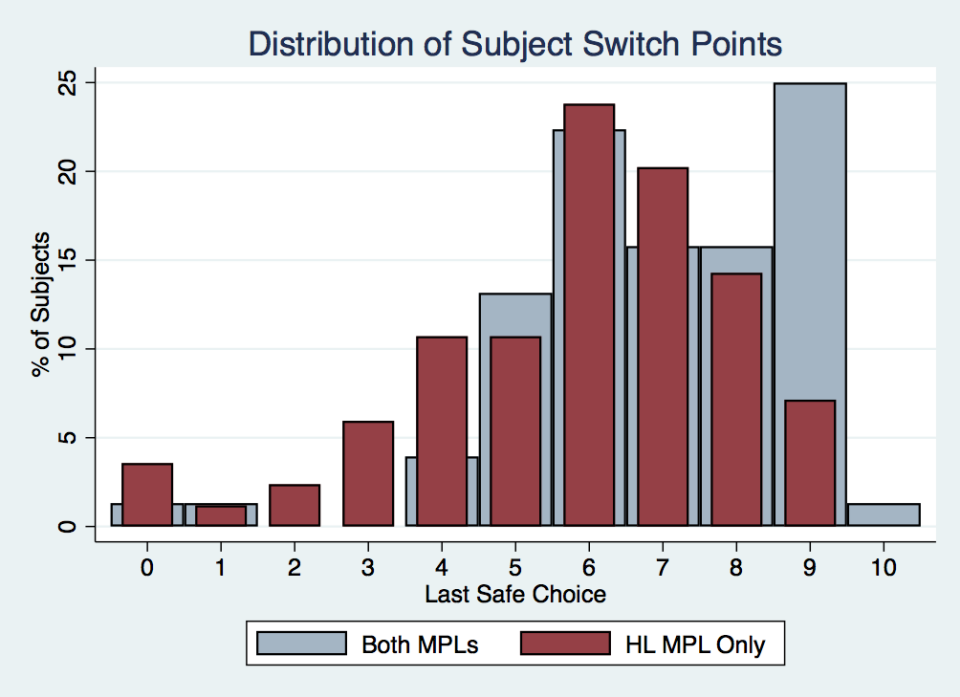


Table 13 includes the results from the regressions of *Last Safe Choice* on the indicated standardized ability score, *Female*, *Economics Major*, a dummy indicator for the HL MPL only treatment, the number of reversals, and earnings in the BCG and TG. In contrast to the results for the BCG and the TG where the coefficient estimates on *Std. Unpaid Score* and the *Std. Paid Score* are very similar across the three specifications, the coefficient estimate on *Std. Paid Score* in Column (2) is more than twice the magnitude of the estimate on *Std. Unpaid Score* in Column (1). However, none of the models generate estimates that significantly differ from zero and they are all positive, which if they were statistically significant would imply that high-ability subjects are more risk averse than low-ability subjects.

The lack of evidence for a inverse relationship between cognitive ability and risk aversion in this study contradicts the findings of a number of studies that conclude individuals with higher cognitive ability tend to be more risk tolerant (see Dohmen

et al. (2018)). However, the studies that find an inverse relationship tend to use self-reports of risk preferences or experiments with low expected payoffs because a random selection procedure is used to determine which subjects will actually be paid for their decisions in the experiment. In contrast, Taylor (2013) finds that the inverse relationship is only significant when the choices are hypothetical, but not when they are real, and Andersson et al. (2016) demonstrates that the increased noise in low-ability subjects choices can make them appear either more risk averse or less risk averse depending upon the MPL used to measure risk preferences.

It is also unlikely that the lack of a significant relationship between cognitive ability and risk preferences is due to peculiarities of the sample because the females in this sample tend to be significantly more risk averse than the males, which is consistent with the conclusion of most studies that have explored the relationship between risk preferences and gender (Croson and Gneezy, 2009; Borghans et al., 2009).

Table 13: HL MPL Risk Aversion Regression Results

Dependent variable is *Last Safe Choice*.			
	(1)	(2)	(3)
Std. Unpaid Score	0.144	—	0.270
	(0.192)		(0.220)
Std. Paid Score	—	0.314	—
		(0.210)	
Std. Dif. Paid-Unpaid	—	—	0.435
			(0.269)
Female	0.623**	0.670**	0.664**
	(0.294)	(0.289)	(0.292)
Economics Major	-0.536	-0.638	-0.630
	(0.427)	(0.430)	(0.442)
H&L MPL Only	-1.061***	-1.011***	-1.010***
	(0.283)	(0.287)	(0.287)
Reversals	1.186***	1.209***	1.203***
	(0.158)	(0.162)	(0.164)
TG Earnings	-0.103*	-0.108**	-0.110**
	(0.053)	(0.052)	(0.052)
BCG Earnings	-0.023	-0.037	-0.034
	(0.060)	(0.061)	(0.061)
Constant	6.711***	6.726***	6.729***
	(0.427)	(0.422)	(0.426)
Observations	158	158	158
Log-likelihood	-312.112	-310.415	-310.116
R-sq.	0.283	0.298	0.301
Wald Tests: Std. Unpaid Score = Std. Paid Score			
Wald F-stat.	—	0.650	0.328
Wald p-value	—	0.422	0.568

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

One subject dropped because of a non-response on *Female*. One subject is dropped because he selected the safe option for H&L Decision #10, which serves as a rationality check on this task.

4 Conclusion

In their search to understand whether and how cognitive ability is related to important economic behaviors, economists have primarily used a protocol that measures ability without incentivizing performance. This contrasts with the standard protocol of incentivizing choices when we measure economic behaviors or preferences because economists typically assess an experiment that offers incentives to be more likely to generate valid measures than an experiment that does not. Yet, there is evidence to suggest that incentives may worsen performance on some tasks (Ariely et al., 2009) or have non-monotonic effects (Gneezy and Rustichini, 2000). In a survey of the literature, Camerer and Hogarth (1999) concluded that incentives tended to improve performance on judgment tasks that required effort. The questions that are often used to measure cognitive ability in experiments and surveys, in particular the cognitive reflection test (CRT) questions, are designed to require effort. Thus, they are the type of task on which incentives are more likely to positively impact subject performance. On the whole, however, the evidence from this experiment suggests incentives do not improve performance on these types of cognitive ability tests, nor do they increase effort as measured by time spent on each question.

This is good news for experimenters on tight budgets: differences in intrinsic motivation are not a significant problem for experimenters who measure cognitive ability with short tests without monetary incentives. Although there are some subjects who appear to be externally motivated and perform better when monetary incentives are offered for correct responses, other subjects perform worse and nearly 80% of the sample score either above the mean on both tests or below the mean on both tests.

Not only did incentives not seem to significantly affect subject performance on the tests, it also does not seem to matter whether the relationship between cognitive ability and several important behaviors is estimated using an incentivized or unincentivized measure of cognitive ability. Consistent with the existing literature, high-ability subjects make lower guesses and earn more money in the beauty contest game, and they tend to be more trusting but not more reciprocating. However, I do not find evidence that higher ability subjects are more risk tolerant, which contradicts the literature that finds an inverse relationship between ability and risk preferences (Dohmen et al., 2018). It is unlikely that the sample size is the cause of this lack of relationship since the estimated coefficient was positive. Moreover, given that the estimated relationships between ability and strategic sophistication, as well as ability

and trust, were consistent with the prior literature, the lack of an inverse relationship between ability and risk aversion seems unlikely to be caused by some peculiarity of the sample. Unless the relationship depends on those at the far lower end of the ability spectrum who are unable to answer basic questions about percentages, which ninety percent of this sample could.

Behavioral economist have demonstrated that introducing monetary incentives can often generate behavior inconsistent with a simple model that does not account for non-economic factors in people's decision making (see Gneezy et al. (2011) for a summary). For example, monetary incentives may crowd out intrinsic motivation or they may result in a state of hyperarousal that causes a decrease in performance (see Camerer and Hogarth (1999) and Ariely et al. (2009) for summaries of this literature). One potential consequence of hyperarousal is the narrowing of attention on fewer dimensions of a problem, which could cause subjects to be less reflective and, thus, perform more poorly on CRT questions. However, the evidence from this study does not indicate subjects significantly increased their effort in response to incentives and, although they do perform slightly worse on some CRT questions, they generally perform the same with or without incentives. It is still possible that the extrinsic motivation induced with incentives offset some intrinsic motivation, but a more likely explanation is that incentives are simply not that important in this context. People who participate in economics experiments in the lab are more likely to have high levels of intrinsic motivation and there is not a significant amount of across-subject variation. Therefore, our conclusions about the relationships between ability and economic behavior that have been obtained with unincentivized measures of ability are probably valid.

References

Andersson, Ola, Jean-Robert Tyran, Erik Wengstroöm, and Hakan J. Holm, "Risk Aversion Relates to Cognitive Ability: Preferences or Noise," *Journal of the European Economic Association*, 2016, *14* (5), 1129–1154.

Ariely, Dan, Uri Gneezy, George Loewenstein, and Nina Mazar, "Large Stakes and Big Mistakes," *The Review of Economics Studies*, April 2009, *76* (2), 451–469.

Baron, Jonathan, Sydney Scott, Katrina Fincher, and S. Emlen Metz, "Why Does the Cognitive Reflective Test (Sometimes) Predict Utilitarian Moral Judgment (and Other Things)?," *Journal of Applied Research in Memory and Cognition*, October 2015, *4*, 265–284.

Berg, Joyce, John Dickhaut, and Kevin McCabe, "Trust, Reciprocity, and Social History," *Games and Economic Behavior*, 1995, *10*, 122–142.

Borgans, Lex, Angela Lee Duckworth, James J. Heckman, and Bas ter Weel, "The Economics of Psychology of Personality Traits," *The Journal of Human Resources*, Fall 2008, *43* (4), 972–1059.

Borghans, L., B.H.H. Golsteyn, J.J. Heckman, and H. Meijers, "Gender Differences in Risk Aversion and Ambiguity Aversion," *Journal of the European Economic Association*, 2009, *7* (2-3), 649–658.

Branas-Garza, Pablo, Praveen Kujal, and Balint Lenkel, "Cognitive Reflection Test: Whom, How, When," *Working Paper*, 2016.

_ , Teresa Garcia-Munoz, and Roberto Hernan-Gonzalez, "Cognitive effort in the Beauty Contest Game," *Journal of Economic Behavior & Organization*, June 2012, *83.*

Branas-Graza, Pablo and John Smith, "Cognitive abilities and economic behavior," *Journal of behavioral and experimental economics*, October 2016, *64*, 1–4.

Buchan, Nancy R., Rachel T. Croson, and Sara Solnick, "Trust and Gender: An Examination of Behavior and Beliefs in the Investment Game," *Journal of Economic Behavior & Organization*, December 2008, *68* (3-4), 466–476.

Burnham, Terence C., David Cesarini, Magnus Johannesson, Paul Lichtenstein, and Björn Wallace, "Higher Cognitive Ability is Associated with Lower Entries in a p-Beauty Contest," *Journal of Economic Behavior and Organization*, 2009, *72* (1), 171–175.

Cagno, Daniela Di and Emanuela Sciubba, "Trust, Trustworthiness and Social Networks: Playing a Trust Game When Networks are Formed in the Lab," *Journal of Economic Behavior & Organization*, August 2010, *75* (2), 156–167.

Camerer, Colin F. and Robin M. Hogarth, "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework," *Journal of Risk and Uncertainty*, 1999, *19* (1-3), 7–42.

Campitelli, G. and M. Labollita, "Correlations of Cognitive Reflection with Judgments and Choices," *Judgment and Decision Making*, 2010, *5* (3), 182–191.

Carl, Noah and Francesco C. Billari, "Generalized Trust and Intelligence in the United States," *PLoS ONE*, 2014, *9* (3), 1–10.

Carpenter, Jeffrey, Michael Graham, and Jesse Wolf, "Cognitive Ability and Strategic Sophistication," *Games and Economic Behavior*, 2013, *80*, 115–130.

Cokely, Edward T. and Colleen M. Kelley, "Cognitive Abilities and Superior Decision Making Under Risk: A Protocol Analysis and Process Model Evaluation," *Judgment and Decision Making*, February 2009, *4* (1), 20–33.

Corgnet, Brice, Antonio M. Espin, Roberto Hernan-Gonzalez, Praveen Kujal, and Stephen Rassenti, "To Trust, or Not to Trust: Cognitive Reflection in Trust Games," *Journal of Behavioral and Experimental Economics*, 2016.

Cox, James C., "How to Identify Trust and Reciprocity," *Games and Economic Behavior*, 2004, *46*, 260–281.

Croson, Rachel and Uri Gneezy, "Gender Differences in Preferences," *Journal of Economic Literature*, 2009, *47* (2), 448–474.

Cueva, Carlos, Inigo Iturbe-Ormaetxe, Esther Mata-Perez, Giovanni Ponti, Marcello Sartarelli, and Haihan Yu, "Cognitive (Ir)reflection: New Experimental Evidence," *Journal of Behavioral and Experimental Economics*, October 2016, *64*, 81–93.

Dave, Chetan, Catherine C. Eckel, Cathleen A. Johnson, and Christian Rojas, "Eliciting Risk Preferences: When is Simple Better?," *Journal of Risk and Uncertainty*, 2010, *41*, 219–243.

Delis, Manthos D. and Nikolaos Mylonidis, "Trust, Happiness, and Households' Financial Decisions," *Journal of Financial Stability*, 2015, *20*, 82–92.

Dittrich, Marcus, "Gender Differences in Trust and Reciprocity: Evidence from a Large-scale Experiment with Heterogenous Subjects," *Applied Economics*, July 2015, *47* (34-36), 3825–3838.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde, "Are Risk Aversion and Impatience Related to Cognitive Ability?," *American Economic Review*, 2010, *100* (3), 1238–1260.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde, "On the Relationship between Cognitive Ability and Risk Preference," *Journal of Economic Perspectives*, Spring 2018, *32* (2), 115–134.

_ , _ , _ , _ , Jurgen Schupp, and Gert G. Wagner, "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences," *Journal of European Economics Association*, June 2011, *9* (3), 522–550.

Drichoutis, Andreas C. and Jayson L. Lusk, "What Can Multiple Price Lists Really Tell Us About Risk Preferences?," *Journal of Risk and Uncertainty*, 2016, *53*, 89–106.

Duckworth, Angela Lee, Patrick D. Quinn, Donald R. Lynam, Rolf Loeber, and Magda Stouthammer-Loeber, "Role of Test Motivation in Intelligence Testing," *Proceedings of the National Academy of Sciences*, May 2011, *108* (19), 7716–7720.

Fehr, Dietmar and Steffen Huck, "Who Knows it is a Game? On Strategic Awareness and Cognitive Ability," *Experimental Economics*, 2016, *19*, 713–726.

Frederick, Shane, "Cognitive Reflection and Decision Making," *Journal of Economic Perspectives*, Autumn 2005, *19* (4), 25–42.

Garbarino, Ellen and Robert Slonim, "The Robustness of Trust and Reciprocity Across a Heterogeneous U.S. Population," *Journal of Economic Behavior & Organization*, 2009, *69*, 226–240.

Gill, David and Victoria Prowse, "Cognitive Ability, Character Skills, and Learning to Play Equilibrium: A Level-k Analysis," *Journal of Political Economy*, December 2016, *124* (6), 1619–1676.

Gneezy, Uri and Aldo Rustichini, "Pay Enough or Don't Pay at All," *The Quarterly Journal of Economics*, August 2000, *115* (3), 791–810.

_ , Stephan Meier, and Pedro Rey-Biel, "When and why incentives (don't) work to modify behavior," *The Journal of Economic Perspectives*, Fall 2011, *25* (4), 191–209.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales, "Does Culture Affect Economic Outcomes," *Journal of Economic Perspectives*, Spring 2006, *20* (2), 23–48.

_ , _ , and _ , "Trusting the Stock Market," *The Journal of Finance*, 2008, *63* (6), 2557–2600.

Harrison, Glenn W., *Hypothetical Bias Over Uncertain Outcomes*, Northampton, MA: Edward Elgar, 2006.

_ and Elisabet Rutström, "Risk Aversion in the Laboratory," in James C. Cox and Glenn W. Harrison, eds., *Research in Experimental Economics*, Vol. 12, JAI Press, 2008, chapter 3, pp. 41–196.

Heckman, James J. and Tim Kautz, "Hard Evidence on Soft Skills," *Labour Economics*, 2012, *19*, 451–464.

Holt, Charles A. and Susan K. Laury, "Risk Aversion and Incentive Effects," *American Economic Review*, 2002, *92* (5), 1644–55.

Keynes, John Maynard, *The General Theory of Employment, Interest and Money* 1936.

Nagel, Rosemarie, "Unraveling in Guessing Games: An Experimental Study," *The American Economic Review*, December 1995, *85* (5), 1313–1326.

Oechssler, J, A Roider, and PW Schmitz, "Cognitive abilities and behavioral biases," *Journal of Economic Behavior & Organization*, 2009, pp. 147–152.

Peters, Ellen, Nathan Dieckmann, Anna Dixon, Judith H. Hibbard, and C.K. Mertz, "Less is More in Presenting Quality Information to Consumers," *Medicaral Care Research Review*, April 2007, *64* (2), 169–190.

Primi, Caterina, Kinga Morsanyi, Francesca Chiesi, Maria Anna Donati, and Jayne Hamilton, "The Development and Testing of a New Version of the Cognitive Reflection Test Applying Item Response Theory," *Journal of Behavioral Decision Making*, June 2016, *29*, 453–469.

Ross, Sheldon, *A First Course in Probability*, 6 ed., Upper Saddle River, New Jersey: Prentice Hall, 2002.

Schwartz, Lisa M., Steven Woloshin, William C. Black, and H. Gilbert Welch, "The Role of Numeracy in Understanding the Benefit of Screening Mammography," *Annals of Internal Medicine*, December 1997, *127* (11), 966–972.

Taylor, Matthew P., "Bias and Brains: Risk Aversion and Cognitive Ability Across Real and Hypothetical Settings," *Journal of Risk and Uncertainty*, 2013, *46*, 299–320.

_ , "Are High-Ability Individuals Really More Tolerant of Risk? A Test of the Relationship Between Risk Aversion and Cognitive Ability," *Journal of Behavioral and Experimental Economics*, 2016, *63*, 136–147.

Thomson, Keela S. and Daniel M. Oppenheimer, "Investigating an Alternate From of the Cognitive Reflection Test," *Judgment and Decision Making*, January 2016, *11* (1), 99–113.

Tu, Qin and Erwin Bulte, "Trust, Market Participation and Economic Outcomes: Evidence from Rural China," *World Development*, 2010, *38* (8), 1179–1190.

Tymula, Agnieszka, Lior A. Rosenberg Belmaker, Belmaker, Amy K. Roy, Lital Ruderman, Kirk Manson, and Paul W. Glimcher, "Adolescents' Risk-taking Behavior is Driven by Tolerance to Ambiguity," *Proceedings of the National Academy of Sciences*, 2012.

Weller, Joshua A., Nathan Dieckmann, Martin Tusler, C.K. Mertz, William J. Burns, and Ellen Peters, "Development and Testing of an Abbreviated Numeracy Scale: A Rasch Analysis Approach," *Journal of Behavioral Decision Making*, April 2013, *26* (2), 198–2013.

5 Appendix

5.1 Conditional Probability Questions

Figure 10: Version A, Q9: MAMMOGRAM and Accompanying Table

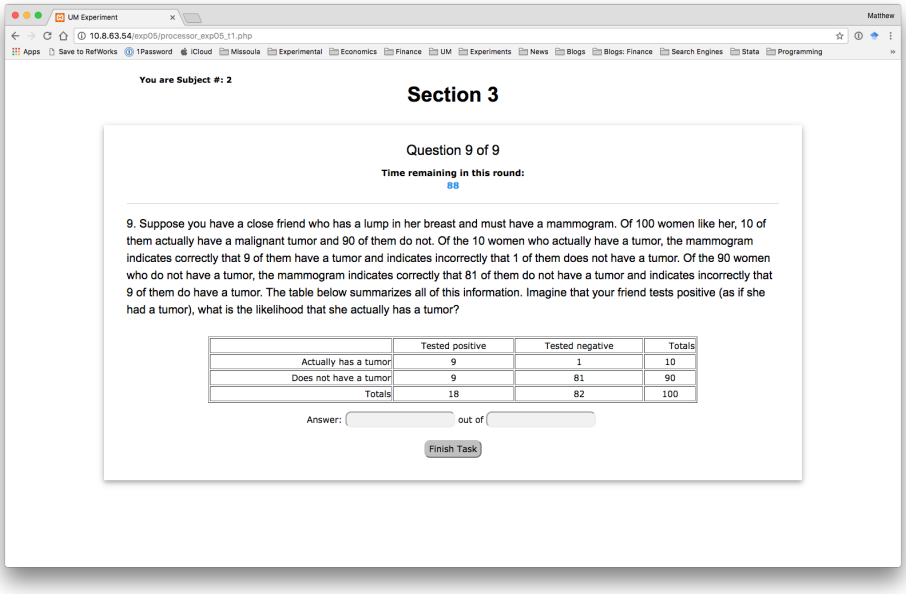
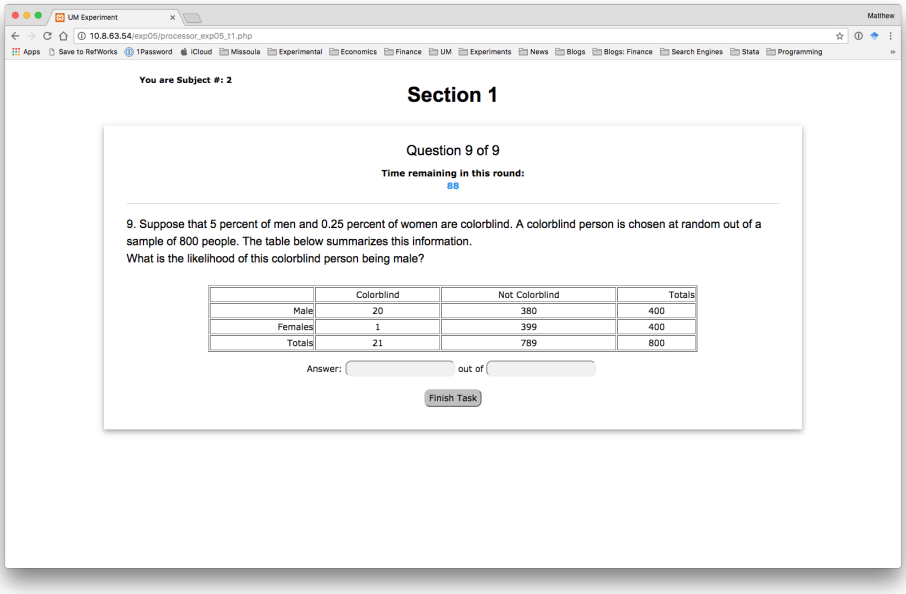


Figure 11: Version B, Q18: COLORBLIND and Accompanying Table



5.2 Ability and Treatment

Table 14: Poisson Results: Ability Scores and Treatment Condition

	(1) Score	(2) CRT	(3) 7-item
main			
Version B	0.016	-0.066	-0.017
	(0.093)	(0.143)	(0.141)
Paid	-0.083	-0.170	-0.148
	(0.093)	(0.135)	(0.134)
Version B × Paid	0.159	0.272	0.272
	(0.126)	(0.198)	(0.190)
Second Test	0.072	-0.057	0.083
	(0.095)	(0.150)	(0.138)
Version B × Second Test	0.074	0.299	0.158
	(0.129)	(0.200)	(0.191)
Paid × Second Test	-0.005	0.067	0.003
	(0.127)	(0.205)	(0.188)
Version B × Paid × Second Test	-0.071	-0.178	-0.127
	(0.088)	(0.142)	(0.127)
Constant	1.548***	1.002***	1.065***
	(0.069)	(0.099)	(0.101)
/			
lnalpha	-3.153***	-1.564***	-1.635***
	(0.747)	(0.519)	(0.500)
Observations	320	320	320
ll	-673.387	-596.102	-619.440
chi2	3670.701	720.891	909.838

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5.3 Beauty Contest Tables

Table 15: Beauty Contest Guesses and Ability Regression Results without Controls

	(1) Round 1	(2) Round 2	(3) Round 3	(4) Round 4	(5) Round 5
Panel A					
Std. Unpaid Score	-4.977***	-6.558***	-4.400***	-5.230***	-4.907***
	(1.852)	(1.475)	(1.472)	(1.719)	(1.549)
Panel B					
Std. Paid Score	-7.071***	-6.254***	-3.646**	-4.511**	-3.900**
	(1.852)	(1.475)	(1.472)	(1.719)	(1.549)
Panel C					
Std. Unpaid Score	-6.866***	-7.301***	-4.585***	-5.551***	-5.019***
	(1.955)	(1.572)	(1.522)	(1.618)	(1.492)
Difference	-7.705***	-3.031	-0.753	-1.309	-0.455
	(2.660)	(2.367)	(2.460)	(2.954)	(2.372)
Observations	160	160	160	160	160

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 16: Beauty Contest Earnings and Ability Regression Results without Controls

	(1) Round 1	(2) Round 2	(3) Round 3	(4) Round 4	(5) Round 5	(6) All Rounds
Std. Unpaid Score	0.155**	0.144**	0.147**	0.053	0.094	0.593***
	(0.073)	(0.064)	(0.068)	(0.068)	(0.070)	(0.176)
Std. Paid Score	0.193**	0.144**	0.126*	0.089	0.062	0.614***
	(0.073)	(0.064)	(0.068)	(0.068)	(0.070)	(0.176)
Std. Unpaid Score	0.198***	0.164**	0.156**	0.081	0.089	0.688***
	(0.075)	(0.072)	(0.074)	(0.075)	(0.073)	(0.187)
Dif. Std. Scores	0.176	0.082	0.034	0.114	-0.020	0.387
	(0.113)	(0.093)	(0.096)	(0.094)	(0.106)	(0.274)
Observations	160	160	160	160	160	160

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5.4 Trust Game Tables

Table 17: Trust Game: Player 1 Tobit Regression Results—Trust

	(1)	(2)	(3)
Player 1 Passed (in dollars)			
Std. Unpaid Score	0.415**	—	0.465**
	(0.192)		(0.208)
Std. Paid Score	—	0.410**	—
		(0.200)	
Std. Dif. Paid-Unpaid	—	—	0.192
			(0.308)
Friend in Session	-0.043	-0.034	-0.043
	(0.532)	(0.533)	(0.531)
Female	-1.282***	-1.284***	-1.265***
	(0.368)	(0.370)	(0.368)
HL MPL Switch Point	-0.046	-0.077	-0.057
	(0.110)	(0.109)	(0.111)
Economics Major	0.151	0.082	0.068
	(0.667)	(0.684)	(0.679)
TG Quiz Attempts	-0.447	-0.323	-0.392
	(0.395)	(0.399)	(0.404)
Constant	4.103***	4.298***	4.154***
	(0.729)	(0.723)	(0.734)
/			
var(e.play1pass)	2.309***	2.326***	2.299***
	(0.439)	(0.442)	(0.436)
Observations	79	79	79
Log-likelihood	-130.068	-130.296	-129.873
Chi-sq.	21.376	20.920	21.765
Wald F-stat.	—	0.001	0.059
Wald p-value	—	0.980	0.809

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 18: Trust Game: Player 1 Tobit Regression Results—CRT

	(1)	(2)	(3)
Player 1 Passed (in dollars)			
Std. Unpaid CRT Score	0.303	—	0.387*
	(0.197)		(0.208)
Std. Paid CRT Score	—	0.376*	—
		(0.194)	
Std. Dif. CRT Paid-Unpaid	—	—	0.344
			(0.291)
Friend in Session	-0.051	0.004	-0.002
	(0.541)	(0.536)	(0.538)
Female	-1.312***	-1.281***	-1.277***
	(0.376)	(0.373)	(0.374)
HL MPL Switch Point	-0.048	-0.091	-0.085
	(0.112)	(0.110)	(0.116)
Economics Major	0.241	0.170	0.159
	(0.681)	(0.678)	(0.682)
TG Quiz Attempts	-0.470	-0.314	-0.329
	(0.403)	(0.402)	(0.416)
Constant	4.143***	4.399***	4.362***
	(0.748)	(0.725)	(0.768)
var(e.play1pass)	2.384***	2.346***	2.345***
	(0.453)	(0.445)	(0.445)
Observations	79	79	79
Log-likelihood	-131.180	-130.494	-130.483
Chi-sq.	19.150	20.523	20.545
Wald F-stat.	—	0.144	0.165
Wald p-value	—	0.706	0.685

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 19: Trust Game: Player 2 Regression Results—Reciprocity & CRT

	(1)	(2)	(3)
Std. Unpaid CRT Score	-0.004	—	-0.008
	(0.029)		(0.030)
Std. Paid CRT Score	—	-0.008	—
		(0.026)	
Std. Dif. Paid-Unpaid	—	—	-0.017
			(0.038)
Player 1 Passed (in dollars)	-0.004	-0.004	-0.004
	(0.024)	(0.024)	(0.024)
Friend in Session	0.092	0.091	0.093
	(0.093)	(0.094)	(0.094)
Female	-0.103*	-0.104*	-0.103*
	(0.056)	(0.057)	(0.057)
HL MPL Switch Point	0.003	0.004	0.004
	(0.009)	(0.009)	(0.009)
Economics Major	-0.047	-0.043	-0.044
	(0.075)	(0.073)	(0.075)
TG Quiz Attempts	-0.062**	-0.062**	-0.060**
	(0.026)	(0.028)	(0.025)
Constant	0.424***	0.422***	0.418***
	(0.129)	(0.129)	(0.130)
Observations	77	77	77
Log-likelihood	15.658	15.697	15.799
R^2	0.097	0.098	0.100
Wald F-stat.	—	0.022	0.014
Wald p-value	—	0.882	0.906

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 20: Trust Game: Regression Results—Earnings

	(1) Player 1 Earnings	(2) Player 2 Earnings
Std. Unpaid CRT Score	0.099	0.194
	(0.274)	(0.534)
Std. Dif. CRT Paid-Unpaid	0.055	0.037
	(0.372)	(0.604)
OP Std. Unpaid CRT Score	-0.020	0.594
	(0.284)	(0.414)
OP Std. Dif. CRT Paid-Unpaid	0.333	0.585
	(0.427)	(0.614)
Friend in Session	-0.585	0.395
	(0.628)	(1.344)
Female	-0.203	-0.149
	(0.540)	(0.908)
HL MPL Switch Point	-0.175	-0.051
	(0.174)	(0.162)
Economics Major	-0.650	-1.034
	(1.206)	(0.972)
TG Quiz Attempts	-0.278	-0.151
	(0.428)	(0.560)
Constant	6.595***	6.359***
	(1.181)	(1.254)
Observations	78	79
Log-likelihood	-163.498	-204.616
R^2	0.060	0.052

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5.5 Risk Aversion Tables

Table 21: CRT and Inconsistency

Logit regressions in which the dependent variable is an indicator variable equal to one if a subject made at least one reversal.			
	(1)	(2)	(3)
Std. Unpaid CRT Score	-0.291*	—	-0.277
	(0.176)		(0.188)
Std. Paid CRT Score	—	-0.190	—
		(0.174)	
Std. Dif. CRT Paid-Unpaid	—	—	0.051
			(0.255)
Female	0.217	0.231	0.222
	(0.353)	(0.354)	(0.356)
Economics Major	-0.474	-0.508	-0.490
	(0.552)	(0.550)	(0.558)
H&L MPL Only	0.240	0.236	0.247
	(0.345)	(0.344)	(0.342)
Constant	-0.899***	-0.892***	-0.903***
	(0.339)	(0.344)	(0.338)
Observations	159	159	159
Log-likelihood	-98.463	-99.235	-98.445
Wald Chi-sq stat.	—	0.336	0.006
Wald p-value	—	0.562	0.938

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 22: Poisson Models of *Last Safe Choice*

	(1)	(2)	(3)
Std. Unpaid Score	0.021	—	0.038
	(0.030)		(0.034)
Std. Paid Score	—	0.043	—
		(0.033)	
Std. Dif. Paid-Unpaid	—	—	0.058
			(0.043)
Female	0.097**	0.104**	0.103**
	(0.047)	(0.047)	(0.047)
Economics Major	-0.082	-0.096	-0.095
	(0.074)	(0.074)	(0.075)
HL MPL Only	-0.170***	-0.161***	-0.160***
	(0.044)	(0.045)	(0.045)
Reversals	0.168***	0.171***	0.169***
	(0.024)	(0.025)	(0.025)
Constant	1.805***	1.797***	1.797***
	(0.049)	(0.050)	(0.050)
Observations	158	158	158
Log-likelihood	-339.238	-338.645	-338.556
Wald Chi-sq. stat.	—	0.437	0.238
Wald p-value	—	0.509	0.626

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 23: Poisson Models of *Last Safe Choice*

	(1)	(2)	(3)
Std. Unpaid CRT Score	0.024	—	0.042
	(0.029)		(0.032)
Std. Paid CRT Score	—	0.047*	—
		(0.029)	
Std. Dif. CRT Paid-Unpaid	—	—	0.061*
			(0.034)
Female	0.090*	0.098**	0.098**
	(0.047)	(0.046)	(0.046)
Economics Major	-0.088	-0.107	-0.106
	(0.074)	(0.074)	(0.075)
H&L MPL Only	-0.178***	-0.169***	-0.168***
	(0.045)	(0.045)	(0.046)
Reversals	0.171***	0.172***	0.170***
	(0.025)	(0.024)	(0.025)
Constant	1.815***	1.808***	1.808***
	(0.048)	(0.049)	(0.049)
Observations	159	159	159
Log-likelihood	-341.730	-341.030	-340.949
Wald Chi-sq stat.	—	0.669	0.333
Wald p-value	—	0.413	0.564

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$