Analysis Plan for AER-RCT Registration

Ceren Bengü Çıbık & Daniel Sgroi

July 16, 2020

The experiment consists of two waves. The first wave was conducted in February, 2020 on Amazon M-Turk. The second wave is going to be conducted in July,2020 on Amazon Mechanical Turk. All participants are given $2 for completing the Human Intelligence Task plus a performance related bonus payment. The experiment uses a between subject design in which subjects are randomly allocated to one of the four groups before they start completing the tasks. The groups only differ in one aspect. Depending on the treatment group subjects are assigned to, they are asked to write about a real-life event in the last 12 months where they were completely honest, or dishonest with various consequences on other people. The waves differ only in terms of the dishonesty task subject face in the last stage of the experiment which will be explained in more detail below.

Experimental Design

The experiment consists of three stages. The first stage includes a questionnaire that aims to collect basic demographic information and to elicit participants' preferences for fairness, risk and integrity and their ethical stance. It also contains a brief version of the Big Five Questionnaire (Rammstedt & John 2007) to measure personality.

In the second stage of the experiment, if the participants are randomly assigned to one of the treatment groups, they are asked to write about an event in their own life. In the Honesty Treatment, they are asked to write about an event involving complete honesty. In the Low Dishonesty Treatment, they are asked to write about an event in which the participant decided not to be completely honest in order to benefit themselves but this dishonesty did not harm anyone else, while in the High Dishonesty Treatment, this dishonesty ended up harming someone else. Subjects in the Control Group directly progress to the third stage of the experiment. We aim to impose different levels of cognitive dissonance to the subjects who are assigned to different treatment groups. We argue that reminding people about their positive self-image would not impose any prior level of cognitive dissonance while reminding them about their negative self-image (dishonest) would impose an initial level of cognitive dissonance which could increase if their dishonesty ended up harming another person.

In the third stage of the experiment, subjects complete dishonesty tasks in a randomized order that aim to elicit dishonest behaviour, two questions to control for the potential demand effect and two questions to elicit their preferences for altruism. The participants are

incentivised by additional potential earnings based on their response in one of the dishonesty tasks which is randomly chosen at the end of the experiment. The dishonesty tasks and the difference among the waves will be discussed in detail below.

Wave 1

In the first wave of the experiment, the dishonesty tasks include the Matrix Puzzle Game (Ariely 2012) and the Cheap Talk Sender Receiver Game (Gneezy 2005). In the matrix task, participants are given 20 different matrices in an image and asked to find two numbers that add up to 10 in these matrices in 5 minutes. Once the time is over, subjects are directed to the next page to report the number of pairs they found. The potential bonus they could make depends on the number of matrices they report they solved. The experiment includes two matrix puzzles where the incentives are \$0.10 and \$0.30 per correct answers. Since the participants' answers are not checked by any examiner, they are free to report any number they wish, but it is not possible to detect dishonesty in an individual level unless they report an infeasibly large number. Therefore, our main comparison is the average number of matrices reported to be solved in the Control Group with the treatment groups to measure dishonesty, though we can also examine infeasibly large reported numbers at the individual level. In the Sender Receiver Game, participants are asked to imagine that they are matched with another anonymous MTurk worker and they need to decide which message to send to their counterparts. There are two possible monetary payments available and are described as 'Option A' and 'Option B'. The choice rests with their counterpart and they can only send either a honest or a dishonest message to their counterparts about the monetary payments associated with each option. The only information their counterparts will have is the signal sent by them. In our design, two different allocations related to Option A and Option B are used but in both of these two cases, Option B always gives higher payoff to the sender than Option A. Therefore, if the player sends a message which states that 'Option B will earn you more money than Option A', it is classified as a dishonest message. Payoff allocations for two tasks are as follows: Option A: \$1 to you and \$X to the other player. Option B: \$X to you and \$1 to the other player where X was set equal to \$1.20 in the low incentive setting, and \$3 in the high incentive setting.

The aim behind our selection of the dishonesty tasks was to vary the psychological cost of lying/cognitive dissonance for the subjects. By asking them to decide to behave dishonestly or honestly under different environments where the dishonesty is more/less costly in terms of the level of cognitive dissonance expected to be experienced, we are able to observe the effect of self-awareness.

We argue that the cognitive dissonance that is expected to be incurred because of a dishonest behaviour in the matrix task will be higher than the cognitive dissonance that is expected to be experienced due to a dishonest behaviour in the Sender Receiver game because of the following reasons. We categorize the games on four characteristics that determine the level of cognitive dissonance/psychological cost if the subjects behave dishonestly. These characteristics are whether the game is a strategic or non-strategic game, whether the final decision is only responsibility of one person, whether the dishonesty is salient or not in the game, and whether the task is ego-relevant or not. We argue that the self-awareness affects the level of dishonesty but the direction of the effect depends on the characteristics of the

games mentioned above.

The Sender Receiver Game. Firstly, if a game has a strategic component the psychological cost of behaving dishonestly would be lower than the case where the one's utility depends on solely his actions. Since this is a game in which you are already working to damage the other person's payoff (in order to boost your own) and having accepted that you are attempting to harm the other player, we might reason that this will reduce the marginal (psychological) costs of lying. In this situation, the main motive of the subjects while deciding for their action is to win the game or gain more payoff than the other person rather than behaving dishonestly or not. Therefore, the decision of behaving dishonestly or the immorality of dishonestly might not be salient in this type of game. We therefore reason that sending a dishonest message in the Sender Receiver Game in our experiment might create zero or small amounts of cognitive dissonance since the players are motivated to win the game and the behaviour of sending a dishonest message might not conflict with their honest self-image because of strategic motives. Also, since the Sender Receiver Game includes lying as a possible action in the list of actions, it is easy to mentally categorize this as selecting a strategy presented to you by the experimenter, rather than undertaking a dishonest action, and so it is easier to overcome cognitive dissonance: we see this as a form of moral "wiggle room" which acts to reduce the salience (and cost) of lying, so again we would expect to see more lying. Moreover, payoffs in the Sender Receiver Game depend on the other player's actions, not only the player himself. This means that the other player is at least partly responsible for her own payoff. If other player follows the signal, then he will suffer the consequences, otherwise he is free to ignore the signal. This can perhaps be interpreted by a player that wishes to lie as meaning that the damage done by lying is at least partly the fault of the other player which reduces the (psychological) cost and again increases the level of dishonesty as compared to the case where the responsibility belongs to only the owner of the dishonest message (like in the matrix puzzle).

The Matrix Puzzle Game. On the other hand, the Matrix Puzzle Game does not have any strategic component and the decision of lying is the responsibility of one person - the subject. Also, unlike the Sender Receiver Game, the subjects do not face a salient option of behaving dishonestly. In addition, the Matrix Puzzle Game is at least partly ego-relevant: if you do well in this task, it is because you are better at the task. Lying about how well you do is therefore lying in two distinct ways: not only lying to boost your payoff, but also lying to yourself about how good you are at the task. All of these components contribute to the fact that dishonest behaviour in the Matrix Puzzle Game is (psychologically) more costly than in the Sender Receiver Game.

Conjecture 1: Twinned with the extra salience of dishonesty generated in the three treatments, we would expect to see a decrease in levels of dishonesty in the Matrix Puzzle Game whereas an increase in the Sender Receiver Game.

Wave 2: The Modified Matrix Puzzle Game

In the second wave of the experiment, we modify the matrix puzzle tasks mentioned above in such a way that the psychological cost of lying is different than the original task.

In the wave 2 version of the Matrix Puzzle Game, as in the wave 1 version of the task participants are asked to report the number of pairs that they find in the matrix puzzle which add up to 10. The only difference from the wave 1 task is the payment scheme. While in the wave 1 task we paid participants per matrix they reported to solve, in the modified version, we pay a lump-sum amount only if they are in the top 50% of the distribution. The lump-sum amount is decided based on the difference between the average payment that participants in the top 50% of the distribution received and the average payment that participants in the bottom 50% received in wave 1 of the experiment. Since our aim is to vary only the psychological cost of dishonesty, we maintain the same average payment in expectation but apply a mean-preserving spread which gives a bonus to those in the top half of the distribution. By changing the payment scheme for this task, we aim to add a strategic component (like in the Sender Receiver Game) to the matrix puzzle to make dishonesty less costly as discussed above. As in the wave 1 task, the experiment includes two versions of this task with low or high material incentive. The lump-sum payments are \$0.72 in the low incentive task and \$2.13 in the high incentive task. This allows us to check that our results are robust to a change in the incentive payments.

Conjecture 2: We expect participants to incur lower levels of cognitive dissonance because of their dishonest behaviour in the wave 2 version of the matrix puzzle than in the wave 1 version, therefore they should behave more dishonestly in the wave 2 version.

Main Hypotheses

Our hypotheses are based on Rabin (1994). In this model, two important factors that affect people's level of dishonestly are relevant to our discussion. These are the material benefits obtained from dishonesty and the psychological cost of dishonesty (or cognitive dissonance)[1] We impose different prior levels of psychological cost of dishonesty with the induction of treatments since a certain level of lying could conflict with one's ideal-self and lead to experience a cognitive dissonance. We assume that there is a threshold where people can still define themselves as an honest person. If a person lies more than this morally acceptable level, they would suffer due to the cognitive dissonance. For simplicity, we assume that if the dishonesty level is less than the threshold, the person will not suffer any cognitive dissonance since they can still see themselves as honest. Rabin (1994) makes two conclusions that are related to our experiment. He suggests that if a person receives lower material utility from engaging in an activity, or feels greater distaste from cognitive dissonance, the lower would be the resulting level of immoral activity that person would undertake.

As discussed in the Experimental Design section, our experiment varies the psychological cost of behaving dishonestly by employing different games. If a dishonest behaviour is more costly in a particular game, we expect to observe less dishonesty. On the other hand, in the

[1]Since in our model the probability of getting caught is zero, we eliminate the material cost of lying.

dishonesty treatment groups, subjects are nudged with a positive level of cognitive dissonance, whereas in the honesty group, it is very small or zero. Therefore, in the dishonesty treatment groups by increasing the starting level of cognitive dissonance, we increase the difference between the individual level of lying and the threshold, so that the cognitive dissonance expected to be incurred after a dishonest act will be higher assuming that updating beliefs about the morality of an action is not possible.

In wave 1 of the study, as in line with our model's prediction, we observed a decrease in the level of dishonesty in the Matrix Puzzle Game and an increase in the level of dishonesty in the Sender Receiver Game when participants were induced to become more self-aware through our treatments.

In the second wave of the experiment, we expect to see a greater reported number of matrices solved on average in the wave 2 version of the Matrix Puzzle Game than in the wave 1 version after the treatment induction. However, at this stage since the extent to which each of the factors discussed above contributes to cognitive dissonance is uncertain, the difference between the level of dishonesty among the modified versions of matrix puzzle and the Sender Receiver game is an empirical question. Our main hypotheses which are supported by our model are as follows.

Hypothesis 1: Self-awareness affects the level of dishonesty. However, the direction is determined by the context under which people make their decision to behave dishonestly or honestly.

Hypothesis 2: If a dishonest behaviour in a task is expected to create a higher (lower) level of cognitive dissonance, regardless of the positive or negative self-awareness imposed, the level of dishonesty will be lower (higher) in this task than in a task where the dishonest behaviour is psychologically less costly.

Combining these two more general hypotheses with our two task-related conjectures produces clear and testable predictions about behaviour related to treatment vs control, across tasks and between wave 1 and wave 2 versions of the tasks.

Data Analysis

Before we start analyzing the data, we will eliminate the subjects who do not satisfy our core treatment induction. This entails eliminating data derived from subjects who produced irrelevant text in the second stage of the experiment. In Wave 1 this accounted for approximately 20% of the sample.

Next we will begin our analysis of the data with an exploration of the descriptive statistics. As a randomization check, we will compare the groups in our experiment by employing the data we collected in the first stage of the experiment.

In order to test our hypotheses, we will start with a mean value comparison test (t-test) between the treatment groups and the control groups for the dishonesty variables obtained from the different tasks mentioned above. We expect to observe an increase in the level of dishonesty in all of the treatment groups in the wave 2 Matrix Puzzle Game as compared to the control group.

After conducting the mean analysis test, we will perform a regression analysis of our main dishonesty variables on the treatment variable. We will include 3 regressions where the first model is the baseline model, the second model includes the set of demographic variables as a control and the third model adds more control variables such as the integrity score, ethic score, the big five personality traits, risk preferences. These analyses will help us to test Hypothesis 1 and begin considering Hypothesis 2. In order to further test Hypothesis 2, we will compare the mean values from the wave 1 and wave 2 versions of the matrix puzzle game.

The next step is to check whether the "moral balancing" argument can be supported by our data. We will use the answers to two Dictator Game questions. After the dishonesty tasks, in both waves of the experiment we ask subjects how much they would like to donate from their potential bonus to *MacMillan Cancer Support* and to give up for researchers to conduct more further sessions in the experiment. We will compare the level of dishonesty in all of the games (across both waves) with the total level of giving. If the moral balancing argument holds, we expect to observe a positive correlation. In addition, even though the dishonesty is not detectable at the individual level for some of our variables, we will compare the dishonest individuals among different dishonesty tasks.

Finally, we will conclude our analysis with an exploratory text analysis. We will use the text data written in the second stage of the experiment to take a closer look at the motives behind the dishonest behaviour.

References

Ariely, D. (2012), *The (Honest) Truth about Dishonesty: How we lie to everyone - especially ourselves*, Harper, New York.

Gneezy, U. (2005), 'Deception: The role of consequences', *American Economic Review* **95(1)**, 384–394.

Rabin, M. (1994), 'Cognitive dissonance and social change', *Journal of Economic Behaviour and Organization* **23**, 177–194.

Rammstedt, B. & John, O. P. (2007), 'Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german', *Journal of Research in Personality* **41**, 203–212.