The Interrogation Game: Using Coercion and Rewards to Elicit Information from Groups

David Blake Johnson
PhD Candidate
The Florida State University
236 Bellamy Building
Tallahassee, FL 32306-2180
djb07c@my.fsu.edu

John Barry Ryan
Assistant Professor
The Florida State University
558 Bellamy Building
Tallahassee, FL 32306-2230
jryan2@fsu.edu

Abstract: In this paper, we examine how interrogators can get potential sources to provide information which entails defecting from their group. In our experiment, subjects are faced with an interrogator either using coercive techniques or offering rewards. We argue that coercion and reward affect individuals who are "conditional defectors" differently. These individuals will defect only when they can justify that selfish action as either fair or truth telling. For subjects who possess the information the interrogator desires, these conditional defectors will provide that information in both treatments because they are simply telling the truth. For ignorant subjects, conditional defectors provide bad information under coercion because honestly stating ignorance leads to unequal outcomes. In the reward treatment, truthfully saying "I don't know" leads to a more equal outcome. This means that interrogators receive more information under coercion, but that information is of lower quality.

Some of the seminal works on group behavior ask this question: how can we encourage individuals to cooperate when it is in their self-interest is to defect from the group? For example, one could model sharing of common property as a Prisoner's Dilemma—individuals have incentives to take all of the public good they can which hurts all members of the group including themselves if everybody acts selfishly (Ostrom 1990). Hence, the key research questions centers on the types of institutions that would encourage cooperation and allow the group to receive larger benefits.

If we consider the Prisoner's Dilemma framed around actual prisoners, the normative consequences are a little murkier. The prisoners are unable to cooperate and as a result are punished because they defect. This is bad for the prisoners, but *good for the society as a whole*. The concentration on how to encourage cooperation has ignored the fact that sometimes people prefer to encourage defection. This is especially true in the case where people want individuals to provide information on members of their group. Examples of this range from a teacher who wants to know what student threw a paper airplane to a police officer attempting to locate where a gang stashes drugs.

Interrogators can incentivize group members to provide information through two means: coercion or rewards. While the use of coercive techniques in interrogation is a highly controversial topic, there is almost no scholarly research on its effectiveness (Conrad and Moore 2010). The difficulties with studying this issue while upholding the standards of ethical research, has led to a reliance on anecdotal and often contradictory evidence (Armshaw 2011). Further, even when scientists do involve themselves in the issue, they often abandon rigor and nuance in the discussion in favor of moral absolutes (Suedfeld 2007).

This paper serves as a first attempt to experimentally determine: (1) how individuals behave under the two types of interrogation and (2) the effectiveness of the two techniques. Ethical concerns obviously prevent research on the harshest methods, but even evaluating the basic hypotheses is difficult if one observes actual interrogation—even in a setting as inconsequential as a high school principal's office. The researcher must know *a priori* the information the source possesses to evaluate the truthfulness of what the source says. As a result, we examine interrogation with an incentivized, group-based experiment. This particular research method is best served at answering a limited set of questions related to the effects of the type of interrogator—rewarding or coercive—the potential source's knowledge state, and the source's desire for a fair outcome as well as his or her desire to tell the truth.

In the experiment, subjects in groups of four are assigned a playing card, but only two of the subjects know what that card is. The computer asks the subjects to reveal the card. If a subject reveals the card, the other group members receive no payment. In the coercion treatment, subjects receive less money than their other group members when they fail to reveal the card. In the reward treatment, the computer provides an additional payoff in exchange for the information.

In all treatments, the individual's dominant strategy is to defect from the group—i.e., to provide information even if that information is potentially inaccurate. We argue that subjects will only defect when they can justify such an action as being something other than selfish. This motivation combines with the treatment to result in very specific predictions. For ignorant subjects, they defect more frequently in the coercion treatment because that is both the selfish action and it results in the more fair outcome. Knowledgeable subjects reveal the card in both situations because it is the selfish thing to do, but truth telling is also the "right" thing to do.

The stakes are higher in the world outside the lab, but the comparison between the treatments should provide some insights. Real world interrogators in the entire range of possible settings offer harsher punishments, but they can also offer greater rewards. We would not suggest that the size of the lab treatment effect is the size of the effect in actual interrogations nor can we speak to effectiveness of extreme coercive techniques that many would consider to be torture. We do believe, however, that we demonstrate that coercive interrogation does not need to be the inevitable behavior of rational interrogators dealing with rational sources as previous models suggest (e.g., Wantchekon and Healy 1999).

Interrogators are Poor Lie Detectors

When an "interrogator"—and by that we mean any individual attempting to obtain information from group members not just a person who holds an equivalent job title—questions an individual, the interrogator is revealing his or her ignorance. The interrogator is not only ignorant about the information about which he or she is asking, but the interrogator does not know if the target under questioning knows the answers either. If interrogators were human lie detectors, the process of questioning would be much easier indeed.

In short, when a source says, "I do not know", the interrogator needs to determine whether or not the source is truly ignorant. If the source is ignorant, then any further interrogation could only result in the source providing bad information. Unfortunately for the interrogator, techniques for determining the truthfulness of sources are not well-developed (Mann, Vrij and Bull 2002). For example, in Vrij et al.'s (2007) experiments, police officers observed interrogators using different techniques to interview mock suspects who had taken place in a staged event. Subsequently, the officers were asked which of the people being

questioned was telling the truth. The officers were more likely to falsely accuse people of lying under harsher questioning techniques and they were more likely to believe that they were accurate in those accusations.

Given the potential issues with coercive techniques that are not controversial, why then is the use of more extreme techniques so prevalent?[1] For example, Parry (2010) details the frequent use of torture after World War II by democracies—or other countries acting on behalf of a democracy—that are signatories to the United Nations Convention Against Torture (CAT). These countries claim that exigent circumstances compel them to use extraordinary means of obtaining information. Democracies are constrained in ways that make the open use of physical coercion difficult. They do, however, practice techniques that physically and mentally stress the potential source without leaving bruises and scars (Ron, 1997; Rejali, 2007).

To understand the prevalence of coercive interrogation, Wantchekon and Healy (1999) model a “game" between an interrogator and a non-cooperative source. In their model, the potential sources have varying levels of resolve and information. They find that once a state selects coercion as a means of educing information they will continue to do so. This occurs because all types of interrogators in their model wish to test the mettle of the victim—i.e., to

[1] Physically coercive techniques, especially instances of torture, likely fall outside the purview of this study. The trauma resulting from such interrogation could cause temporary memory loss making a knowledgeable source temporarily ignorant. A most extreme case of this comes from the memoirs of Paul Aussaresses, a chief intelligence officer during the Battle for Algiers (Arrigo 2004). Aussaresses reports that his use of torture during the interrogations of detainees assisted in ceasing the Algerian insurgency and bombings. There were instances, however, where the torture was so extreme that the detainees died before revealing their secrets.

determine what type of source they have. Unless all potential sources are of the "strong" type—that is, they will never reveal information—then all interrogators will use and continue to use some level of coercive interrogation.

Wantchekon and Healy (1999) "assume that the state has the means to verify the truthfulness of the information provided by the victim" (600). Without this assumption, states may be less willing to use coercive techniques. We argue that both knowledgeable and ignorant sources can provide inaccurate information and the interrogator is unable to tell the difference at the time of the interrogation. If states were to act on that information, then that could prove quite costly. For example, knocking down one door when a bomb is behind another one does not prevent the attack, it embarrasses the state, and punishes the innocents whose door the state destroyed.

The Truth and What Is Fair

In determining how individuals will behave under different forms of interrogation, there are three individual level characteristics that must be considered: (1) the individual's knowledge state—i.e., do they know the information the interrogator desires or not; (2) the individual's regard for a fair outcome relative to his or her group; (3) the individual's aversion to lying. Outside of the lab, an interrogator does not know the knowledge state of the interrogation target, but knowledge state is easily manipulated in an experiment. The other two factors are underlying personality characteristics and potentially context dependent. For example, when a teacher asks a student who pulled a prank the student may have no trouble lying to the teacher, but may reveal the truth when his or her mother serves as the interrogator—or vice versa depending on the student's relationship with the two interrogators.

The nature of the two forms of interrogation—coercive or rewarding—means that in manipulating the type, the experimenter is also manipulating the fairness of the outcomes. Individuals may experience disutility when they receive a smaller payout than other members of their group, but, at the same time, their utility may also decrease when they receive a greater payout than other members of the group (Fehr and Schmidt 1999).[2] For individuals who know the information the interrogator desires, telling the truth increases the inequality of payoffs regardless of the interrogation treatment and to the benefit of sources who provide the information—they either avoid the punishment themselves or they receive a reward.

For ignorant individuals, telling the truth and revealing their ignorance has different effects on the equality of payoffs. Under coercion stating that you are ignorant will result in punishment for only the individual and an unfair outcome in which the individual suffers. The ignorant individual should reveal false information in the hopes that the interrogator will spare the full punishment. Under reward, the ignorant individual could guess at some information in

[2] Bolton and Ockenfels (1998) propose an alternative model to Fehr and Schmidt (1999) in that individuals' maximize a motivation function; holding individual earnings fixed, utility is maximized when the average payoff is relatively close to their own. Although these theories generally result in similar predications, certain payoff configurations could lead to differences (Engelmann and Strobel, 2004). For example, in Bolton and Ockenfels a subject is indifferent between all group members receiving equal payoffs and the subject earning the average amount in which some subjects receive a large payoff and others a small payoff. In Fehr and Schmidt, the subject would prefer that all group members receive the same payoff. Our argument is motivated by Fehr and Schmidt's conception of inequality aversion, but there is no reason to believe that switching to Bolton and Ockenfels' model would result in different expectations.

an attempt to receive at least part of the reward. This results in an unfair payoff in which the individual benefits. The fairer outcome is for the individual to truthfully state that they do not know.

Hence, if an individual does not possess information, we should expect them to reveal information (i.e., lie) regardless of their level of lie aversion under coercion than reward. Knowledgeable informants, however, should not be any more likely to reveal the truth in either condition. This leads to two seemingly contradictory hypotheses:

H_1. Individuals will be more likely to reveal information under coercion than reward.

H_2. The quality of information will be lower under coercion than reward.

The seemingly paradoxical nature of these two hypotheses is the result of more ignorant individuals revealing information under coercion. They provide information to the interrogator, but most of the time the information is not of use to the interrogator.

In some interrogation situations, the source will not want to lie to the interrogator either because of some institution which punishes lying specifically (Lupia and McCubbins 1998) or because the individual is averse to lying in general or is concerned about the harm the lie will do to the other parties (Gneezy 2005). Hurkens and Kartik (2008) argue that there are two types of individuals: (1) people who will never lie and (2) people who will lie whenever it would be beneficial to do so regardless of how the effects other group members. Hence, there are some individuals who will not even tell "Pareto white lies"—that is, a lie that will benefit everyone (Erat and Gneezy 2012).

In the case of knowledgeable subjects under interrogation, telling the truth and revealing information is a selfish action—it increases their payoffs while lowering the payoffs of the other group members. This is true under both interrogation techniques. Hence, sources could be of

two types similar to the types identified by Hurkens and Kartik (2008): (1) subjects who always tell the truth; (2) subjects who will tell the truth when doing so benefits them.[3] These individuals will reveal information because truth telling will lead to higher payoffs.

H_3. Knowledgeable subjects will take the higher payoff when doing so is explicitly framed as truth telling.

For both ignorant and knowledgeable individuals, the same logic can explain their actions. There is a set of individuals who are "conditional defectors" who will reveal information depending on the context. What is required to get these individuals to speak is a justification for their actions other than simply wanting the highest possible payoff. For ignorant subjects, this comes under coercive interrogation because seeking the highest payoff is also the fair action. For knowledge subjects, as long as the interrogation is about revealing factual information, the source can take the higher payoff.

Experimental Design

Comparison to Previous Experiments We call our experiment "The Interrogation Game". In the experiment, subjects are placed in groups playing against an outside party who requires a piece of information in order to take an action. Only some of the group members have the information. To ascertain the piece of information, the outside party "interrogates" one randomly chosen

[3] One can see this behavior manifest itself in the behavior of some whistleblowers or people who write tell-all books. It is possible these people are telling the truth because they believe telling the truth is the "right" thing to do. It is also possible these people are telling the truth to put themselves in the limelight. In that case, it appears they are doing something altruistic, but, in fact, they are doing it for selfish reasons.

group member, who may or may not know the required information. The outside party takes an action based upon what that source says. [4] The outside party may punish sources for not revealing the information or reward them for revealing accurate information. That is, unlike most models of group behavior in which 3rd parties punish or reward to encourage cooperation (e.g., Fehr and Fischbacher, 2004), the 3rd party in this experiment encourages defection among the group members.[5] If the source provides accurate information, the other group members receive a payoff of zero.

The experiment that we present has elements that are similar to two other experiments. First, there is the Volunteer's Dilemma (Diekmann, 1985). In the Volunteer's dilemma, N players are grouped together and given the choice to cooperate at a cost of *K* or defect which is costless. If at least one group member selects to cooperate, the subject(s) who choose to cooperate earn *U-K*, while the group members who defect earn *U*. If they all defect, the subjects earn nothing. Hence, the dilemma is whether someone is willing to pay a cost for all members to receive a benefit.

[4] Often in the outside world, the interrogator has priors based upon the detainee's individual history (e.g., the principal has some knowledge of the student's credibility and the police interrogator is aware of the detainee's criminal background). Our computer "interrogator" acts upon a commonly known rule thus removing a strategic element of the game and allowing us to investigate truth telling without confounding effects from strategic interaction.

[5] There are exceptions. Apesteguia, Dufwenberg and Selten (2007) examine corporate leniency where an outside party encourages defection from a cartel. Our experiment differs in that we introduce players who do not have the information the outside party requires and have no strategic interaction with other players.

Rapoport's (1988) Volunteer's Dilemma is framed specifically around a coercive interrogation. Subjects are told they are to imagine that they are in a prison or boarding school and the authorities want to find out who committed some transgression. If no one confesses all subjects are punished severely, but if one subject confesses the other group members would not be punished while the one that confessed would only be mildly punished. Rapoport finds that between 40 and 50 percent of subjects cooperate—that is, they are willing to bear the cost on behalf of the group—regardless of whether or not the nature of the social dilemma has been explained to the subjects. Our experiment differs in that subjects do not choose whether or not to volunteer. The chosen source has been "volunteered" and in the coercion treatment must decide whether they are willing to pay a cost to ensure the best payoff for other group members.

For this reason, the experiment also bears similarity to the Dictator Game (for a review see List, 2007) where one player is chosen to divide an endowment between him or herself and another player. Like the Dictator Game, what we call The Interrogation Game is not a game in any strict sense because there is no intersection between player behaviors and payoffs. The potential source in The Interrogation Game is similar to the dictator in that he or she sets the payoffs for all players. In most cases, dictators choose to keep the entire endowment for themselves, but substantial positive offers are common.[6]

[6] This depends greatly on the conditions of the experiment. Hoffman, McCabe, and Smith (1996) find that as anonymity is reduced, zero offers become less common. However, it is possible that the steps taken to insure anonymity also result in subjects becoming skeptical that their offers will actually reach their partners prompting more selfish behavior (Bolton, Katok and Zwick, 1998).

Subjects and Groups We recruited subjects for our group-based experiment at a public university in the southern United States. For each session, we recruited sixteen student subjects using the ORSEE recruitment system (Greiner, 2004). Upon being randomly seated, subjects were given instructions and anonymously assigned to groups of four.[7] Subjects then either participated in a coercion treatment or a reward treatment for ten periods—we will explain the treatments in the following section. Following those ten periods, subjects were assigned to new groups and participated in the other treatment for ten periods.

After completing these main experimental treatments subjects participated in a modified Dictator Game and a Holt and Laury (2002) risk aversion task. Subjects participated in these final two experiments primarily to increase their earnings considering that some subjects will receive zero earnings in both of the main treatments. Each session lasted approximately 60 minutes. Subjects earned $6.79 on average in the main experimental treatments in addition to the $10 show up fee and whatever they earned in the other tasks.

In all, 160 subjects participated in ten sessions. In half of the sessions, subjects were randomly assigned to their groups of four. After the first ten periods, they were re-matched so that they did not play with the same subjects as they did in the previous 10 periods. In the other sessions, subjects were assigned to groups using a standard minimal group paradigm (MGP) procedure (Tajfel 1970). Subjects were placed into groups based on their preferences for different abstract paintings.[8] Chen and Li (2009) find that when subjects are put into groups

[7] The exact instructions can be found in the Supporting Information.

[8] Prior to the first treatment, subjects were divided into groups based on their preferences between paintings by Paul Klee and Wassily Kandinsky. For the second treatment, subjects were divided into groups based on their preferences between paintings by Koko the gorilla and Congo

based upon their preference for paintings subjects behave more pro-socially toward those within their own group in comparison to groups where group assignment was done randomly. As we find no statistically significant differences between the two types of group assignment, we present only the pooled results here.[9] The results broken down by group type are available in the Supporting Information. In all cases, subjects were accurately told that their group members in one treatment may not be their group members in the other treatment.

Three of the sessions with random groups began with subjects participating in the coercion treatment for ten periods and then in the reward treatment for ten periods. In the remaining two random group sessions, subjects participated in the reward treatment for 10 periods before being re-matched to participate in the coercion treatment. Similarly, in the MGP group sessions, three began with the coercion treatment and two began with the reward treatment.

---

the chimpanzee. Like subjects' preferences for artwork by Klee and Kandinsky, we have no reason to believe preferences for either primate is correlated with other subject characteristics. Subjects were not told who the painters were and their groups were named simply "Group A", "Group B", "Group C", or "Group D". Subjects were told they were placed into those groups because they "liked the same paintings."

[9] Why did the MGP treatment fail to generate differences in behavior? One possibility is that the instructions read to the randomly assigned groups already activate the minimal group identity. Specifically, subjects are told that they "are playing as a group against the computer." Hence, even though, the assignment to groups differed in the two treatments, both may result in minimal groups.

The Treatments At the start of each experimental round, each group is randomly assigned a playing card—one of the four aces.[10] The subjects are told that this is their group's card and that they are playing against the computer which is trying to guess the card. Half of the group members see the playing card while the other half of the group simply sees the text "unknown". Subjects are randomly assigned to see the card or to see the text "unknown" in each period—subjects, therefore, could be ignorant in one period and knowledgeable about the card in another.

All subjects are then prompted by the computer to reveal their group's card. Subjects have the option of revealing a card, with no requirement of truthfully revealing the card, or to indicate that they do not know the card. The computer randomly then selects one of the group members to "listen to" and payoffs are assigned based upon the randomly selected subject's decision. It is important to note that, since the "interrogator" is a computer, it is not a peer of the subjects. It is clearly an outside party the way it would be in interrogations outside of the lab.

If subjects have seen the card, there are three possible actions: (1) truthfully reveal the card, (2) reveal an incorrect card, or (3) claim to not know the card. However, if a subject does not know the card, there are only 2 possible actions: (1) truthfully state that they do not know the card or (2) reveal a card at random. If a subject does not know the card and selects to reveal a random card, there is a one in four chance the subject will randomly select the correct card.

Table 1 displays the payoffs the subjects receive for each of the different possible decisions in the two treatments.[11] In the coercion treatment, the subject chosen to act as the

[10] Screen shots of the experiment are available in the Supporting Information. The experiment was programmed using zTree—software for running economic experiments (Fischbacher 2007).

[11] Subjects are paid based on two randomly chosen periods—one in each treatment. Subjects do not know their actual earnings until after both treatments are complete.

source earns $5 if they reveal the correct card and the other group members receive nothing. The source receives nothing if he or she says, "I don't know" and only $2.50 if he or she tells the computer the wrong card. In both cases, the other group members receive $5.

In the reward treatment, the source receives $10 if he or she reveals the correct card and $7.50 if he or she reveals an incorrect card.[12] The payoffs for the other group members are the same in both treatments.

While all subjects are asked to make a decision, the computer randomly chooses to "listen to" one subject. The computer then "guesses" a card based on the information obtained from the chosen subject. The computer reveals whether or not it correctly guessed the card—if the chosen subject claimed to not know, the computer displays that it guessed incorrectly. Subjects do not know who was chosen or the decisions of their fellow subjects; they only know the outcome of the period. Subjects participate in each treatment for ten periods. All of the analyses conducted will cluster standard errors to account for the repeated observations on the same subjects embedded in experimental sessions.

Receiving Accurate (and Inaccurate) Information

We begin the analysis by testing the hypotheses related to the information the interrogator—that is, the computer in the experiment—will receive. Recall that H_1 states the

[12] Subjects in the reward treatment receive $7.50 even for accurate information to keep the two treatments parallel and because the subjects are behaving as if they are trying to defect from the group. In fact, ignorant subjects who provide information are defecting from the group since they run the risk of sticking group members with the zero payoff in order to gain more money for themselves.

interrogator will receive more information under coercion and H_2 states the information will be of lower quality under coercion. Figure 1 displays the proportion of the time that the computer could have used the information obtained from the subjects to correctly guess the card compared to how frequently the computer would have guess the wrong card or been unable to guess a card. It is important to remember that sometimes when subjects revealed the correct card, they were guessing what the card is because they did not know it. They happened to be correct, but they did not have any more information than the computer.

Subjects revealed the correct card 48.4% of the time in the coercion condition compared to 44.8% in the reward condition.[13] This difference is not quite statistically significant at the .05 level in an *F*-test of equal proportions with clustered standard errors (F=3.47 with 150 degrees of freedom; p=0.06). There is much stronger support for hypotheses H_2. Subjects were much more likely to reveal an inaccurate card in the coercion condition (F=11.80; p=0.00). Subjects revealed the wrong card 43.8% of the time in the coercion condition, but were 7.1 percentage points less likely to reveal an incorrect card in the reward condition. The difference is largely the result of how much more frequently subjects clicked the "I don't know" button in the reward condition (F=39.58; p=0.00). In the punishment condition, subjects said they would take the zero payoff for claiming not to know the card only 7.8% of the time. In the reward condition, subjects chose to take $5 and said that they did not know 18.6% of the time.

Subjects are less likely to reveal an incorrect card in the reward condition as we predict, but Figure 1 does not explain why we see this result. Knowledgeable subjects in the reward

[13] Remember, all subjects made a decision. Only one subject decision was randomly chosen to set the payoffs. The results, therefore, show the decisions all subjects made and not just the chosen subject.

condition can receive an extra $2.50 by lying about what the card is. This would increase the inequality in payments, but would not harm the earnings of other group members. Thus, one might expect knowledgeable subjects to reveal incorrect cards with greater frequency in the reward condition. On the other hand, ignorant subjects might not want to accidentally harm their fellow group members in the reward condition by guessing the correct card when they are only trying to gain an extra $2.50.

To explain the results in Figure 1 and to check for over-period effects, Figure 2 displays the frequency of the decisions by period for knowledgeable and ignorant subjects in each treatment. The figure also divides the subjects by whether their session participated in the coercion or reward treatment first. The decisions for subjects who participated in a treatment first are display periods 1 through 10, while the decisions for subjects who participated in a treatment second are displayed in periods 11 through 20.

The first thing to notice is that there do not appear to be any over-period trends in the data. The main factors affecting behavior are the subject's knowledge state and the treatment. The top graphs display the results for knowledgeable subjects, first in the coercion treatment then in reward treatment. The figure shows that knowledgeable subjects did not change their behavior on the basis of the treatment. In both treatments, subjects revealed the correct card about 75% of the time, the wrong card about 22% of the time, and claimed to not know about 3% of the time. It would appear on the basis of these results that subjects took on one of two types when they were knowledgeable. Most maximized their payoffs, but some took the greatest benefit they could without harming the other group members.

When subjects were ignorant, however, the treatment had an effect in support of H_2. In the coercion treatment, ignorant subjects truthfully stated that they did not know what the card is

only about 12.5% of the time. In the reward treatment, ignorant subjects honestly stated that they did not know what the card is 32.5% of the time. By revealing a card, ignorant subjects are running the risk of harming the other group members. In the coercion treatment, however, it is the "fair" thing to do. If an ignorant subject does not reveal any card, they receive nothing in coercion while the group members receive $5 with certainty. If they do reveal a card, their expected payoff is $3.13 and the expected payoff for the other group members is $3.75. Since the payoffs are more equal if ignorant subjects reveal a card in coercion, ignorant subjects should reveal a card regardless of their level of inequality aversion. In reward, everyone will have an equal payoff if subjects do not reveal a card. Thus, in reward, ignorant subjects who care about equality should honestly state that they do not know anything.

There does appear to be an order effect occurring among the ignorant subjects. Ignorant subjects who participate in the reward treatment first state that they did not know what the card is 40% of the time in that treatment. If the reward treatment is second, then they choose "don't know" only 27% of the time. Table 2 presents a logit model of the decision making of ignorant subjects with fixed effects to account for the repeated measures on subjects. In the model, the dependent variable is coded one if the subject truthfully chose "I don't know" and zero if the subject guessed a card. The table shows that ignorant subjects chose "I don't know" more frequently in reward condition regardless of which treatment they participated in first, but that the effect was larger if they participated in the reward treatment first. In a test that both coefficients in the model are equal, χ^2=7:52 with 1 degree of freedom and p=0.01.

These results suggest that subjects who participated in the coercion treatment first got into the habit of revealing a card potentially lowering the payoffs of their fellow group members. Because of this order effect, we are potentially underestimating the difference in how much

inaccurate information the third party receives between the treatments. If we only look at the first treatment that subjects participated in, overall subjects reveal the wrong card 44.1% of the time in coercion and 35.6% in reward. This is a slightly larger difference than the difference shown in Figure 1.

Who Tells the Truth?

The preferred outcome for an interrogator is for the source to tell the truth. That is, the interrogator wants a source who has the information to reveal it and the interrogator wants a source who does not know to say, "I don't know". The first probit model in Table 3 predicts who tells the truth based on the treatment and the subject's knowledge state. The second model includes an interaction between these two variables. We also present a model with three control variables. First, we include a period counter to account for potential period effects. We also include one period lags on the subject's behavior and whether the computer guessed the card correctly. These variables control for, respectively, the potential stickiness of subject behavior and for whether subjects change their behavior based on how they believe other group members are behaving.[14]

The dependent variable in the probit models in Table 3 is coded one if the subject told the truth and zero if the subject lied. Knowledgeable subjects told the truth if they revealed the correct card and ignorant subjects told the truth if they clicked the "don't know" button. The first

[14] The Supporting Information includes a logit model which replicates the first and second models with fixed effects for subjects. These models more directly control for subjects who tend to exhibit the same behavior. The substantive results are the same as in the probit models in Table 3.

model includes only the treatment and knowledge state variables. Both variables are statistically significant and in the expected direction. As expected, subjects tell the truth less frequently under coercion and ignorant subjects are less likely to reveal their ignorance. When we include the interaction effect in the second model, however, we see that the effect of the treatment on truth telling only holds for the ignorant subjects.

This result holds when we include other control variables in the third model.[15] The positive, statistically significant coefficient for *Previous Truth Teller* suggests that subjects' behavior is sticky, though, the statistically significant effect of *Period* shows subjects tell the truth less frequently in later periods. The lack of a statistically significant effect for *Previous Correct Card* demonstrates that subjects do not react to the outcome. It does not appear that they reciprocated the behavior of the other subjects.

The results from the models confirm the problem with coercion that we observed in the previous analyses. Even if we ignore all ethical concerns with interrogators using coercion, coercion is suboptimal because it increases the risk of receiving false information from sources with no actual information. False statements from ignorant sources are often suspected of making coercion an ineffective mechanism for educing information (e.g., Blakeley 2007). This is clearly the case in our laboratory experiment. Of course, we cannot state precisely how the results would change if coercion was the use of force rather than a monetary loss, but there is anecdotal evidence that force often induces false statements from ignorant sources (Armshaw 2011).

Self-Serving Lie Aversion

[15] The N size is smaller in the third model because we drop the first period of each treatment due to the fact that there is no lag for the first period.

To this point, we have shown that ignorant subjects are more likely to tell the truth in the reward condition because honestly revealing ignorance is unfair to them. We have yet to demonstrate that the knowledgeable subjects used the fact that they were telling the truth to justify their selfish behavior. For this reason, we replicate the experiment removing the possibility for subjects to lie by removing the experiment's frame. Half the group members had three choices (BLUE, TEAL, and GREEN) and other half of the group had two (PURPLE, and GREEN). Each of these choices had equivalent payoffs to subjects participating in the original framed experiment. Specifically, GREEN's payoff is identical to that of a subject selecting "Don't Know"; BLUE is equivalent to a subject truthfully revealing the card; TEAL is equivalent to a subject lying about the card. Last, PURPLE pays BLUE 25 % of the time and TEAL the other 75 % of the time.

The left side of Figure 3 shows the decisions made by subjects in the framed experiment by treatment and knowledge state. The right side displays the decisions made by subjects in the unframed version of the experiment. As before, the subjects who can set payoffs with certainty—the subjects who knew the card in the framed treatment—are not effected by treatment. They make the same decisions regardless of whether they are choosing within the coercion payoff scheme or the reward payoff scheme. Subjects who have uncertain payoffs are affected by the treatment and in the same manner as in the framed experiment. It is unfair of them to take the lowest payoff under the coercion scheme. Hence, they only take the lowest payoff 16% of the time. In the reward scheme, taking the lowest payoff keeps the payoffs equal while inequality increases if they attempt to earn more money. For this reason, the percent of subjects taking the lowest payoff doubles.

The frame affects the decisions only of those people who can set payoffs. In the framed condition, subjects take the highest payoff leaving the other group members with the lowest possible payoff nearly 75% of the time. They can justify this action as not selfish, but acting truthfully—to respond with a different card or "I don't know" would be a lie. In the unframed version, this justification is removed. Now, taking the largest payoff is simply a selfish decision. Accordingly, subjects who can set the payoffs are much less likely to take the highest payoff in the unframed version. Instead more subjects take the TEAL option in the unframed version. This is the equivalent of revealing the wrong card in the framed version. The subject takes a lower individual payoff, but the group ends up with the social optimal outcome. In fact, the percentage of subjects taking the lowest payoff doubles in the unframed version. This is very strong evidence that in the unframed version subjects were unwilling to take an action that would clearly be labeled as selfish.

Alternate Explanations and External Validity

We have argued that a combination of concerns for fairness and selfish truth telling explain the pattern of behavior observed by subjects in our experiment. A potential alternative explanation for the decisions would rely on prospect theory (Kahneman and Tversky 1979) or simply loss aversion for the knowledgeable subjects since there is no uncertainty for those subjects. In this case, the argument would state that the subjects react differently under coercion because they do not want to take a loss in payoffs. If this is the case, then we would not need to consider concerns about the fairness of the outcome for themselves or other group members.

There are several problems with this explanation. First, it relies on an assumption regarding the reference point with which subjects compare their final payoffs. Subjects may

believe that $5 is the reference point because that is the payoff everyone receives if the card is not revealed. In that case, receiving $2.50 or nothing under coercion is a loss. Alternatively, the subjects never leave the lab with less money than they entered. Further, the instructions never state that the subjects begin with an endowment of $5. In this case, the subjects would view no money as the reference point. Hence, they would never be on the loss side of the ledger. They would still have to cope with missing out on potential gains. These foregone gains do not seem to have the same psychological burden as actual losses (Kahneman, Knetsch, and Thaler 1986).

More importantly, loss aversion alone cannot explain the actions of the knowledgeable subjects or the subjects who can act as dictator in the unframed experiment. There is no treatment effect for these subjects at all. If they simply wanted to avoid losses, they would take the higher payoffs in the unframed version of the coercion treatment, but in the reward treatment. The results suggest there are two types of subjects. First, some subset of subjects will take the largest payoff regardless of fairness or truth telling. The other subset will sometimes take the larger payoff, but only when they can find a way to justify it.

This typology of subjects informs how one should think about the external validity of the experiment. Much like the "conditional cooperation" observed in public goods games (Fischbacher, Gächter, and Fehr 2001), the subjects who are affected by the treatment were "conditional defectors". Hence, the proportion of conditional defectors in the population outside the lab will affect the proportions of behaviors observed. If a group is made up almost entirely of the type that will never take the selfish payoff—potentially like a committed terrorist cell—then neither technique would reveal information. If a group is made up almost entirely of the selfish type—potentially like an elementary school class—the both techniques would reveal information. It remains the case however, that rewarding is the weakly dominant strategy

because it does not result in conditional defectors revealing false information while coercion does.

Still, we must proceed with caution concerning the external validity of all results. In interrogation settings outside the lab, whether in a principal's office or a police station, the incentives are often much larger. The disincentives to remain quiet when facing a coercive interrogator are much higher than the risk of losing five dollars. However, the rewards that an interrogator can offer are similarly much larger and go beyond simple monetary compensation for information—in an extreme example, the state can offer to protect the family members of a potential source from a war torn country. Thus, while everything in actual interrogations is scaled up, the comparison between the two types of incentives should be similar. Behavior in Dictator Games and Ultimate Games, for example, is remarkably stable even when the monetary stakes are raised to levels greater than weekly wages (Slonim and Roth 1998). Further experiments with higher stakes are necessary to see whether the size of the treatment effect is conditional on the size of the stakes.

We would say that this study probably cannot shed light on interrogations involving physical coercion. In part, this is because it is unclear how physical trauma would affect the behavior of a target. Just as importantly, our experiment assumes that the state uses coercion to obtain information. These results do not speak at all to states that use torture to stifle opposition or during ethnic strife (Poe and Tate 1994; Parry 2010). In those cases, the visible effects of torture are required to demonstrate ruthlessness (Rejali 2007).

Conclusion

This paper argues that individuals certain individuals will take a selfish action only when they can justify that as being the "right thing to do" in some way. In this case of our experiment, that justification comes from subjects selfishly telling the truth if they are knowledgeable. Further, ignorant subjects lie because lying decreases inequality in the coercion condition even though it is a selfish act.

These behaviors suggest that reward is a better interrogation technique than coercion. This is because the "conditional defectors" behave exactly the same way in both treatments when they are knowledgeable, but there is a significant treatment effect when they are ignorant. When they are ignorant they provide poor information to avoid a low payoff under coercion. Under reward, they truthfully state that they do not know because any other action would be purely selfish. Our experiment provides evidence of the common suspicious that coercion is an ineffective means of interrogation because of false statements from ignorant sources (e.g., Blakeley 2007).

This paper also contributes to the literature on how subjects behave in experiments. The extant research has shown that punishment is an effective mechanism to increase cooperation (e.g., Fehr and Gächter 2002). Our experiment centers on instances in which society wants encourage individuals to defect from a group and we show reasons why rewards might be a preferable treatment. Further, we put a spin on the previous literature on lie aversion. That literature argues many individuals will lie only when doing so benefits them (Hurkens and Kartik 2008). In our experiment, we show that you can incentivize people to tell the truth. Instead of the selfish lies observed in previous studies (Erat and Gneezy 2012), our subjects participate in selfish truth telling.

References

Armshaw, Patrick. 2011. The Torturer's Dilemma: Analyzing the Logic of Torture for Information. PhD thesis Florida State University.

Arrigo, Jean Maria. 2004. “A Utilitarian Argument Against Torture Interrogation of Terrorists." *Science and Engineering Ethics* 10(3):543 - 572.

Blakeley, Ruth. 2007. “Why Torture?" *Review of International Studies* 33(3):373 - 394.

Bolton, Gary E. and Axel Ockenfels. 1998. “ERC: A Theory of Equity, Reciprocity, and Competition." *The American Economic Review* 90(1):166 - 193.

Bolton, Gary E., Elena Katok and Rami Zwick. 1998. “Dictator Game Giving: Rules of Fairness Versus Acts of Kindness." *International Journal of Game Theory* 27(2):269 - 299.

Chen, Yan and Xin Li. 2009. “Group Identity and Social Preferences." *The American Economic Review* 99(1):431 - 457.

Conrad, Courtenay Ryals and Will H. Moore. 2010. “What Stops the Torture?" *American Journal of Political Science* 54(2):459 - 476.

Diekmann, Andreas. 1985. “Volunteer's Dilemma." *The Journal of Conflict Resolution* 29(4):605 - 610.

Engelmann, Dirk and Martin Strobel. 2004. “Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments." *The American Economic Review* 94(4):857 - 869.

Erat, Sanjiv, and Gneezy, Uri. 2012. "White lies." *Management Science* 58(4): 723-733.

Fehr, Ernst and Klaus M. Schmidt. 1999. “A Theory of Fairness, Competition, and Cooperation." *The Quarterly Journal of Economics* 114(3):817 - 868.

Fehr, Ernst and Urs Fischbacher. 2004. “Third-party Punishment and Social Norms." *Evolution and Human Behavior* 25(2):6387.

Fehr, Ernst and Simon Gächter. 2002. “Altruistic Punishment in Humans.” *Nature*, 415:137-140.

Fischbacher, Urs. 2007. “z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics* 10(2):171 - 178.

Fischbacher, Urs, Simon Gächter and Ernst Fehr. 2001. “Are people conditionally cooperative? Evidence from a public goods experiment.” *Economic Letters*, 71: 397-404.

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

Greiner, Ben. 2004. An Online Recruitment System for Economic Experiments. In *Forschung und wissenschaftliches Rechnen*, ed. Kurt Kremer and Volker Macho. Gottingen, Germany: Datenverarbeitung pp. 79 - 93.

Hoffman, Elizabeth, Kevin McCabe and Vernon L. Smith. 1996. “Social Distance and Other Regarding Behavior in Dictator Games." *The American Economic Review* 86(3):653 - 660.

Holt, Charles A. and Susan K. Laury. 2002. “Risk Aversion and Incentive Effects Quick View." *The American Economic Review* 92(5):1644 - 1655.

Hurkens, Sjaak, and Navin Kartik. 2009. "Would I lie to you? On social preferences and lying aversion." *Experimental Economics* 12.2: 180-192.

Kahneman, Daniel and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47(2):263 - 292.

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler. 1986. “Fairness as a Constraint on Profit Seeking: Entitlements in the Market.” *The American Economic Review,* 76(4): 728-741.

List, John A. 2007. “On the Interpretation of Giving in Dictator Games." *Journal of Political Economy* 115(3):482 - 493.

Mann, Samantha, Aldert Vrij and Ray Bull. 2002. “Suspects, Lies, and Videotape: An Analysis of Authentic High-Stake Liars." *Law and Human Behavior* 26(3):365376.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. New York: Cambridge University Press.

Parry, John T. 2010. *Understanding Torture*. Ann Arbor, MI: University of Michigan Press.

Poe, Steven C. and C. Neal Tate. 1994. “Repression of Human Rights to Personal Integrity in the 1980s: A Global Analysis." *The American Political Science Review* 88(4):853 - 872.

Rapoport, Anatol. 1988. “Experiments with N-Person Social Traps I: Prisoner's Dilemma, Weak Prisoner's Dilemma, Volunteer's Dilemma, and Largest Number." *The Journal of Conflict Resolution* 32(3):457-472.

Rejali, Darius. 2007. *Torture and Democracy*. Princeton, NJ: Princeton University Press.

Ron, James. 1997. “Varying Methods of State Violence." *International Organization* 51(2):275-300.

Suedfeld, Peter. 2007. “Torture, Interrogation, Security, and Psychology: Absolutistic versus Complex Thinking." *Analyses of Social Issues and Public Policy* 7(1):5563.

Tajfel, Henri. 1970. “Experiments in Intergroup Discrimination." Scientific American 223:96-102.

Vrij, Aldert, Samantha Mann, Susanne Kristen and Ronald P. Fisher. 2007. “Cues to Deception and Ability to Detect Lies as a Function of Police Interview Styles." *Law and Human Behavior* 31(5):499 - 518.

Wantchekon, Leonard and Andrew Healy. 1999. “The Game of Torture." *The Journal of Conict Resolution* 43(5):596 - 609.

Table 1. Payoffs by treatment for the subject the computer chooses and other group members.

	Coercion Treatment	Reward Treatment	Others in Both
Accurate Card	500	1000	0
Inaccurate Card	250	750	500
"Don't Know"	0	500	500
Expected Value of Ignorant Subject Guessing	312.5	812.5	375

Table 2. Did the order of the treatments affect the behavior of ignorant subjects?

	Coef.	Z-Value
Reward First	2.074	8.01
Reward Second	1.181	5.97
N (Subjects)	1253 (126)	
AIC	859.021	

The dependent variable is coded one if the ignorant subject chose "I don't know" and zero if the subject chose to guess a card. The excluded category is the punishment treatment. Logit model with fixed effects for repeated measures on subjects—34 subjects are dropped because of no variance in the dependent variable.

Table 3. Predicting subject truth telling.

	Coef.	Z-Value	Coef.	Z-Value	Coef.	Z-Value
Coercion	-0.312	-4.99	0.035	0.4	0.102	1.16
Ignorant	-1.407	-15.94	-1.067	-10.32	-1.078	-10.06
Coercion*Ignorant	--------		-0.738	-5.39	-0.769	-5.31
Previous Truth Teller	--------		--------		0.518	7.59
Previous Correct Card	--------		--------		0.026	0.49
Period	--------		--------		-0.022	-2.35
Constant	0.797	10.29	0.620	7.91	0.472	4.46
N (Subjects)	3200 (160)		3200 (160)		2880 (160)	
A.I.C.	3516.156		3462.43		3038.994	

Probit models with standard errors adjusted for clustering on subjects.

Figure 1. Proportion of subject decisions by treatment. Error bars represent 95% confidence intervals.

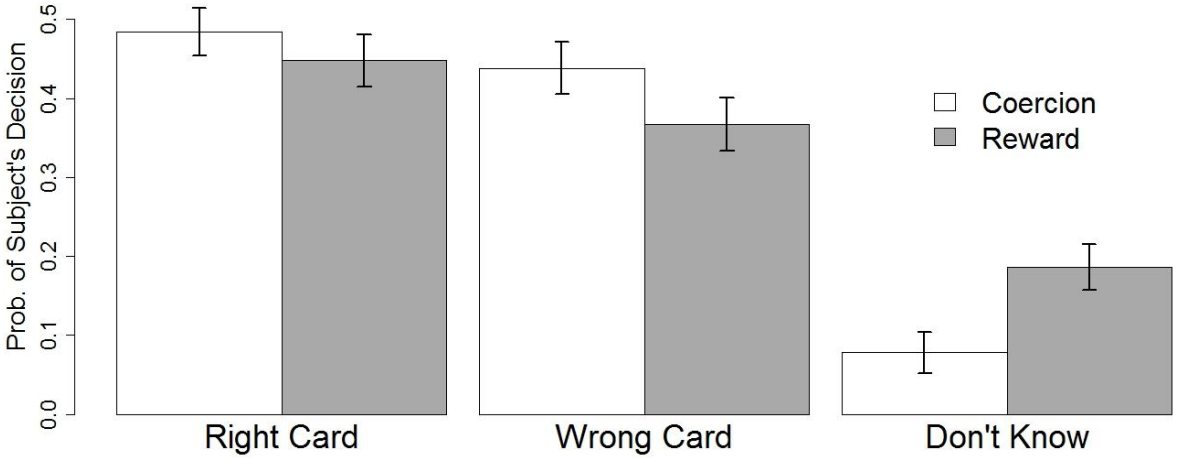


Figure 2. Subject decisions by period for ignorant and knowledgeable subjects.

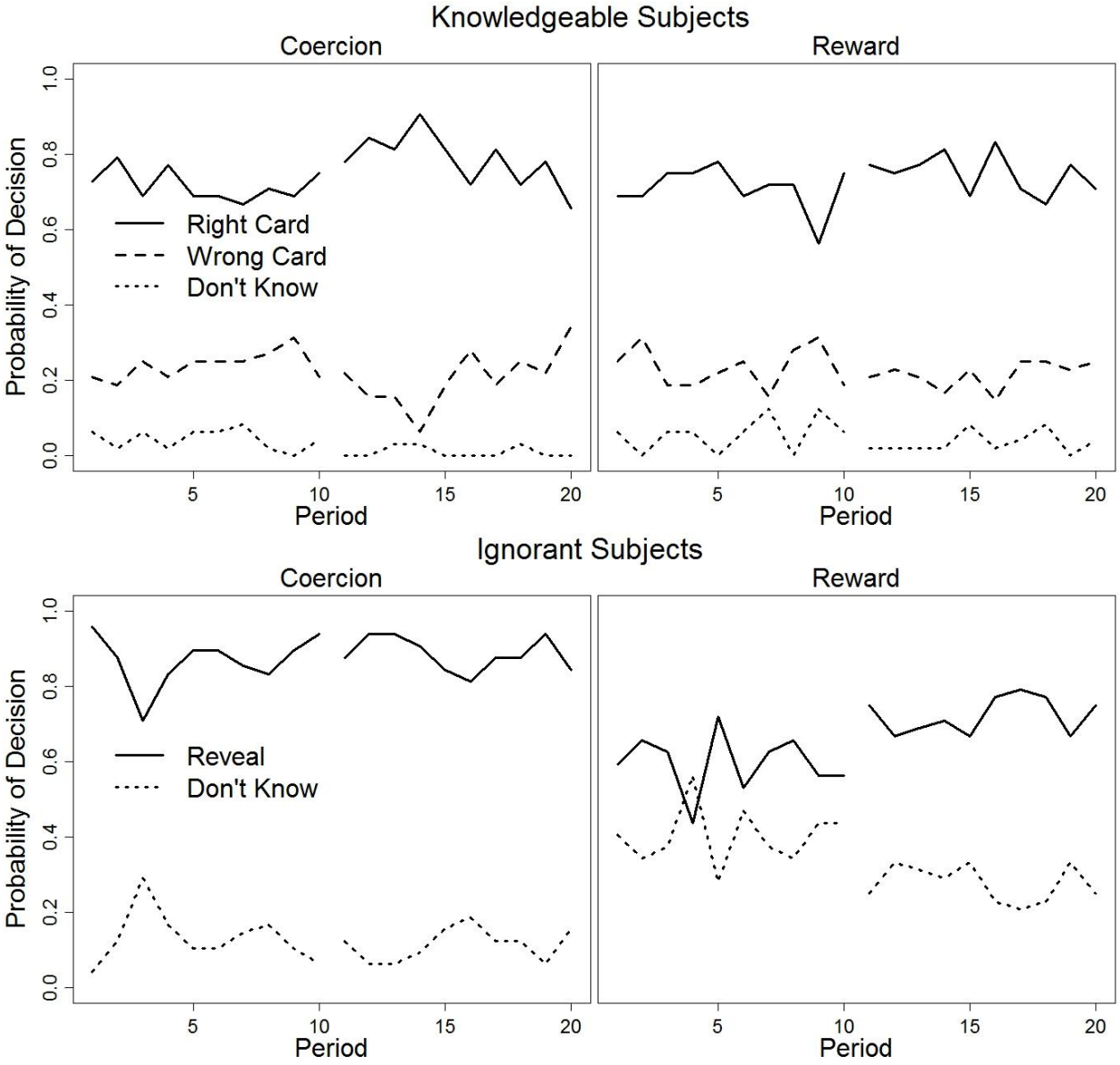
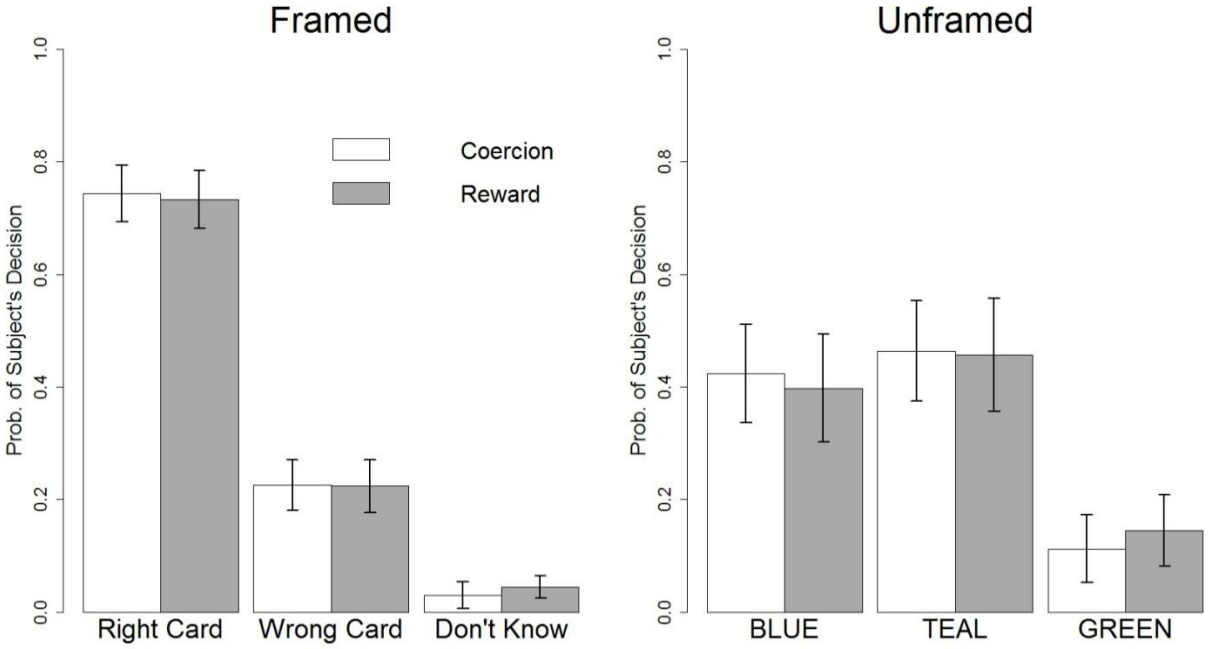


Figure 3. Comparing behaviors in the framed and unframed version of the experiment. Error bars represent 95% confidence intervals.

A. Subjects who can set the payoffs (Knowledgeable subjects in the framed version)



B. Subjects with uncertain payoffs (Ignorant subjects in the framed version)