Do You Reward and Punish In The Way You Think Others Expect You To?

Omar Al-Ubaydli and Min Sok Lee

June 2009

Discussion Paper

Interdisciplinary Center for Economic Science
4400 University Drive, MSN 1B2, Fairfax, VA 22030
Tel: +1-703-993-4850 Fax: +1-703-993-4851
ICES Website: www.ices-gmu.org
ICES RePEc Archive Online at: http://edirc.repec.org/data/icgmuus.html

Do you reward and punish in the way you think others expect you to?[1]

Omar Al-Ubaydli and Min Sok Lee[2]

June 2009

Abstract

This paper addresses three questions: (1) When deciding on whether to reward or punish someone, how does how you think others expect you to behave affect your decision? (2) Does it depend upon whether others expect you to reward them vs. punish them? (3) What is the interpretation of such a causal effect? We investigate these questions using a modification of the lost wallet trust game (Dufwenberg and Gneezy (2000)) that permits punishment. Like previous studies, we collect data on what second-movers think that first-movers expect them to do by directly eliciting the second-movers' expectations. Unlike previous studies, we ensure exogeneity of these expectations by instrumenting for them. The instrument is the expectations of neutral observers which are disclosed to second-movers prior to the elicitation of second-movers' expectations. We find that what you think others expect you to do has a zero causal effect on both reward and punishment decisions. We also find that it is important to instrument for second-order expectations because they are endogenous. We interpret these findings in terms of models of guilt-aversion and intentional reciprocity.

JEL codes: D63, D64, D84

Keywords: Behavioral confirmation, guilt, intentions, reciprocity, reward, punishment.

[1] We wish to thank Martin Dufwenberg, Uri Gneezy, Dan Houser, John List and Nat Wilcox for helpful comments, and Jason Aimone for help in running the experiment.

[2] Al-Ubaydli (corresponding author): Department of Economics and Mercatus Center, George Mason University, 4400 University Drive MSN 3G4 Fairfax, VA 22030 USA. Email: omar@omar.ec. Tel: +1-703-993-4538. Fax: +1-703-993-1133. Lee: Citadel Group Foundation, Chicago, IL.

1. Introduction

Reward and punishment are critical to regulating economic relationships, even in one-shot settings. Investigating their determinants – especially the potentially controllable ones – is an important step towards understanding optimal incentive schemes. One such malleable determinant is what you think others expect you to do. In this paper, we address three questions:

1. When deciding on whether to reward or punish someone, how does how you think others expect you to behave affect your decision?
2. Does it depend upon whether others expect you to reward them vs. punish them?
3. What is the interpretation of such a causal effect?

Previous studies include the lost wallet version of the trust game (Dufwenberg and Gneezy (2000), Berg et al. (1995)), shown in figure 1.

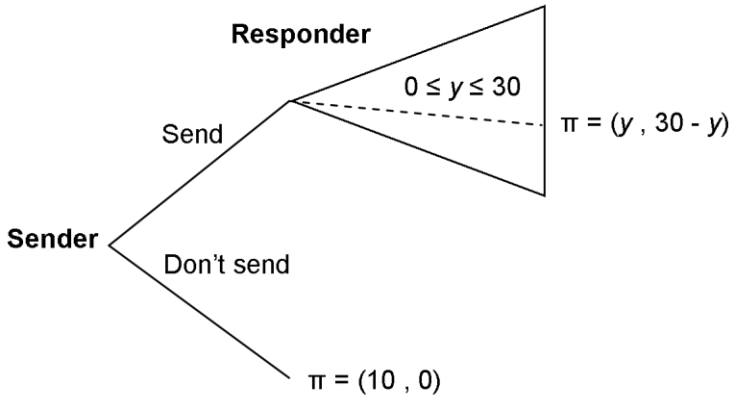

Figure 1: The lost wallet game (Dufwenberg and Gneezy (2000))

The sender starts with $10 and the responder starts with nothing. The sender can either keep the $10, ending the game, or she can send the $10 to the responder. If the sender decides on sending the $10, they are tripled, and the responder unilaterally decides how much of the $30 to return to the sender (y).

Let y' denote the sender's expectation of y. Let y'' denote the responder's expectation of y', referred to as the responder's second-order expectation of y. In other words, y'' is what the responder thinks that the sender expects the responder to send back. By eliciting y'' in simple variants of the trust game, Dufwenberg and Gneezy (2000) and other studies find a positive relationship between y'' and y.[3] This is an example of *behavioral confirmation*: you are more likely to behave in a certain way if you think that others expect you to behave in that way. This is to be contrasted with its obverse: *behavioral disconfirmation*.

[3] Guerra and Zizzo (2004), Charness and Dufwenberg (2006), Bacharach et al. (2007).

The aforementioned studies interpret this instance of behavioral confirmation as representing guilt-aversion: when y'' exceeds y, the responder is failing to fulfill what she believes to be the sender's expectations. The responder feels guilty about disappointing the sender, and will feel guiltier the larger the difference between y'' and y. A guilt-averse responder will therefore send back more money when she thinks that the sender expects more back.

In the trust game, the responder's only alternative to a materialistic best response is to reward the sender.[4] What if the responder is also allowed to punish the sender? In punishment decisions, do we expect behavioral confirmation or disconfirmation? We here consider an extended form of the trust game that allows for punishment.

In addition to extending the scope from reward to punishment decisions, we also investigate the interpretation of the data. In particular, there are theories of behavioral disconfirmation, such as the model of intentional reciprocity.[5] Observing behavioral confirmation is evidence in favor of guilt-aversion. However we explore why observing behavioral confirmation is not necessarily evidence against models that predict behavioral disconfirmation, and, crucially, why observing neither (which we find) is not necessarily evidence against guilt-aversion.

A final issue that we explore is how second-order expectations are observed. Dufwenberg and Gneezy (2000) elicit second-order expectations directly from the responders.[6] Since this treatment variable is not randomly induced by the experimenter, the design risks endogeneity bias.[7] To estimate the endogeneity bias, we transmit the expectations of non-playing observers to responders and then elicit the second-order expectations of responders. We then use the expectations of the observers as instruments for the second-order expectations of the responders.[8]

Our results are as follows. When using observer expectations as an instrument for second-order expectations, second-order expectations have no effect on both reward and punishment behavior, i.e., we find neither behavioral confirmation nor behavioral disconfirmation. When we elicit expectations directly, we find behavioral confirmation in both reward and punishment decisions, suggesting that elicited expectations are endogenous.

[4] Rewards are deviations from best responses that increase the sender's payoff. Punishments are deviations from best responses that decreases the sender's payoff.

[5] The intentional reciprocity model explains reward decisions as resulting from a desire to reciprocate kind actions, and punishment decisions as resulting from a desire to reciprocate unkind actions. See Schopler and Thompson (1968), Pruitt (1968), Tesser et al. (1968), Greenberg and Frisch (1972), Blount (1995), McCullough et al. (2001) and Ames et al. (2004) as well as Geanakoplos et al. (1989), Rabin (1993), McCabe et al. (2003), Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006) and Cox et al. (2007).

[6] As do the other studies in fn 3. In our study, these are incentivized by using a quadratic scoring rule and the elicited expectations of the senders. The senders' expectations are also incentivized by using a quadratic scoring rule and the behavior of the responders.

[7] Charness and Dufwenberg (2006) refer to this possibility as the false-consensus effect on p1594.

[8] Ellingsen et al. (2009) and Reuben et al. (2008) use an alternative design: elicit the senders' expectations and then report them to the responders. We discuss their design in section 2B.

The remainder of this paper is organized as follows. Section 2 is the experimental design. Section 3 is the empirical results. Section 4 is the discussion. Section 5 is the summary and conclusion.

2. Experimental design

A. The judgment game

The judgment game is shown in figure 2. It is a non-linear, continuous version of the games in Dufwenberg and Gneezy (2000) and Offerman (2002). It is a variant of the trust game where the responder can punish as well as reward.

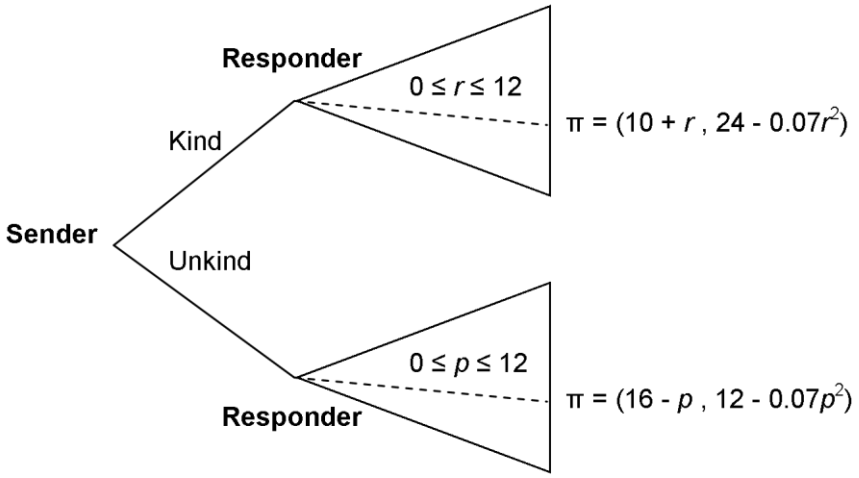

Figure 2: The judgment game

The sender starts with \$16 and the responder with \$12. If the sender plays $unkind$, the payoffs are unchanged. The responder can then punish the sender by reducing the sender's payoff by $p \in [0,12]$. This costs the responder $0.07p^2$.[9]

If the sender plays $kind$, then she transfers \$6 to the responder, which are then doubled. The responder can then reward the sender by increasing the sender's payoff by $r \in [0,12]$. This costs the responder $0.07r^2$.

The quadratic cost of reward (punishment) means that the marginal cost of reward (punishment) rises from \$0 to \$1.7 as r goes from 0 to 12 (p goes from 0 to 12). We selected a quadratic cost as it implies an interior solution under conventional models of behavioral preferences.[10]

[9] In the experiment, we used a neutral frame. Senders were $Blues$ and responders were $Reds$. $Kind$ was $Dash$ and $unkind$ was $Solid$. Reward was $Increase\ Blue's\ earnings$ and punish was $Decrease\ Blue's\ earnings$. Also, payoff consequences of actions were not expressed as the sender starting with an amount that she can choose to send to the responder. Rather the sender was choosing between two actions each with a certain payoff consequence. See the instructions in the appendix.

Let r'' be the responder's second-order expectation of r, i.e., what value of r she thinks that the sender expects her to pick. Similarly, let p'' be the responder's second-order expectation of p.

Research question 1: What is the relationship between reward and second-order expectations of reward, $(\partial r/\partial r'')$? What is the relationship between punishment and second-order expectations of punishment, $(\partial p/\partial p'')$?

Research question 2: Do $\partial r/\partial r''$ and $\partial p/\partial p''$ differ?

Research question 3: Do the answers to research questions 1 and 2 depend upon how data on r'' and p'' are collected?

We detail the alternative data collection methods in the next section.

B. Procedure

Subjects were recruited by email using a campus database at George Mason University. Each session had 14-20 subjects. Senders and responders were in different rooms and roles were assigned randomly. There was no communication and each sender was anonymously matched with a unique responder. Subjects were paid in private.[11]

Given the comparative complexity of the judgment game's payoffs, subjects were given a diagram of the payoffs (see the instructions in Appendix 3). They were also required to complete a short quiz to confirm their ability to locate payoffs on the diagram.

Responders selected their move using the strategy method. While the interchangeability of hot and cold decisions remains an open empirical question, we follow existing studies in using it.[12]

In the elicitation treatment, we collected data on (r'', p'') by direct elicitation: responders were asked to state their second-order expectations. They were incentivized using a quadratic scoring rule; to reward their accuracy, we also collected data on senders' predictions of (r, p), which we denote (r', p').[13] Expectations were elicited after action choices. To test if the act of choosing had any effect on the reported expectations, we ran sessions with observers whose sole task was to form second-order expectations (we had 56 observations from observers). Kolmogorov-Smirnov and Mann-Whitney tests failed to reject the hypothesis that the distribution of (r'', p'') was the same for actual responders vs. observers (all p-values exceeded 40%).

Direct elicitation may imply that second-order expectations are endogenous. To guarantee exogeneity, in the instrumental treatment we collected data in a similar manner to Bacharach et al. (2007), but

[10] Especially Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006).
[11] Average earnings were $14 for a session that lasted about an hour. We used an exchange rate to transform experimental dollars into US dollars.
[12] Charness and Dufwenberg (2006), and implicitly Bacharach et al. (2007) and Guerra and Zizzo (2004).
[13] To explain the quadratic scoring rule, we followed the wording in Offerman et al. (1996).

analyzed the data in a novel way.[14] Before directly eliciting the second-order expectations of each responder, we showed them the average expectations of 10 senders from previous sessions $(\bar{r}', \bar{p}')$. These senders were essentially acting as pseudo-observers in subsequent sessions. The responders were told that the senders whose expectations were being transmitted were being rewarded for their accuracy, and that those senders' expectations were never disclosed to those senders' partners. After directly eliciting (r'', p''), we could use $(\bar{r}', \bar{p}')$ as an instrument for (r'', p'').

In an effort to maximize power, we picked the highest and lowest values for each element of the reported averages $(\bar{r}', \bar{p}')$, yielding four combinations.[15] The instructions did not state that the 10 senders whose expectations were averaged were randomly selected.

Finally, to compare our results with Ellingsen et al. (2009) and Reuben et al. (2008), we replicated their design in the inducement treatment. After eliciting (r', p') (the senders' expectation of (r, p)), we transmitted each sender's expectations to her responder partner. To avoid strategic manipulation of reported expectations, senders were not told that their expectations would be transmitted to responders, and this was made known to responders.

One could make the argument that it is deceptive because the responders feel that we are deceiving the senders' good faith by transmitting their expectations without the senders' knowledge (for more on deception, see Gneezy (2005)). In other words, we are withholding information from subjects that they might like to know. If the responders do regard this as deceptive, then this may affect their subsequent behavior and therefore imply a loss of experimenter control. Nevertheless comparing the results of this treatment with those of the instrumental treatment sheds light on whether any such potential deception effect exists, in addition to serving simply as a check of whether the results are robust to different methods of inducing variation in second-order expectations.[16]

3. Empirical results

We ran 13 sessions during spring 2008, yielding 78 observations from elicitation treatments, 74 from instrumental treatments and 76 from inducement treatments (plus 56 from observers). In each session, we had to throw away several observations due to errors by subjects. When we asked subjects to make predictions, rather than asking them for their (second-order) predictions of (r, p) directly, we asked

[14] We actually also use Bacharach et al.'s (2007) data. See result 1 below.

[15] The high and low values of $\bar{r}'$ were 1 and 7, respectively. The high and low values of $\bar{p}'$ were 1 and 8, respectively.

[16] This raises another potential issue. There is the issue of negative externalities on the rest of the profession. This is a gray area, but we believe that our design is not deceptive. A very similar design is used by Al-Ubaydli and Lee (2009). Fershtman and Gneezy (2001) and Holm and Nystedt (2005) are also examples of designs where information procured from one set of subjects is delivered to another set without the explicit knowledge of the first set.

them for their predictions of the payoffs of responders and senders.[17] Some of the subjects gave inconsistent predictions, e.g., predicting that both partners would have a payoff of 12 points conditional on the sender playing $kind$, while others gave nonsensical ones, i.e., payoffs outside the feasible range. We therefore omit such observations from the results.

Result 1: In the judgment game, $\partial r/\partial r'' = 0$ and $\partial p/\partial p'' = 0$: both reward and punishment fail to exhibit behavioral confirmation or behavioral disconfirmation.

In model 1 in table 1, we instrument for elicited second-order expectations using the average first-order expectations of 10 observers.[18] The estimated causal effect of second-order expectations is positive but statistically insignificant in both reward (p-value = 29%) and punishment (p-value = 51%) choices.

	Model 1	**Model 2**	**Model 3**
Treatment	Instrumental	Elicitation	Inducement
Constant	1.93	0.51	3.05**
Standard error	(2.28)	(0.79)	(1.23)
Kind dummy	-0.42	0.22	-0.76
Standard error	(3.58)	(1.07)	(1.58)
r'' x Kind dummy = [A]	0.60	0.42***	0.21
Standard error	(0.57)	(0.13)	(0.20)
p'' x (1 - Kind dummy) = [B]	0.46	0.68***	0.13
Standard error	(0.69)	(0.19)	(0.21)
P-value of Wald test of [A] = [B]	0.87	0.27	0.79
Observations	66	59	57
R^2	0.08	0.31	0.08

Table 1: Regression results

The dependent variable in all models is actual reward (r) or punishment (p). 'Kind dummy' is a dummy variable taking the value 1 when the sender has played $kind$. (r'', p'') denote elicited (or induced in the inducement treatment) second-order expectations. In the instrumental treatment, the results are from a 2SLS regression where the instrument is the average expectations $(\bar{p}', \bar{r}')$ of 10 other participants. In the inducement treatment, second-order expectations were induced by reporting the expectations (r', p') of the senders. Asterices denote statistical significance (* = 10%, ** = 5%, *** = 1%).

[17] We asked them to report their expectations in payoff space because that data was used in a different experiment.

[18] The first-stage regression (omitted for parsimony and available upon request) confirms that the observer first-order expectations are indeed positively correlated with the elicited second-order expectations. In the reward choice, the estimated correlation is 0.46 (p-value < 1%) and in the punishment choice it is 0.38 (p-value < 3%).

It is reasonable to suggest that insufficient data (rather than a zero causal effect) may be driving the statistical insignificance. To explore this, recall that our design mimics that of Bacharach et al. (2007). They did not estimate the causal effect of second-order expectations using instrumental methods (one of our contributions), but this can still be done using their data.[19] In Appendix 1, we do this and find that, in a dataset with 160 observations, the causal effect of second-order expectations on reward choices is negative and statistically insignificant. This is consistent with result 1.

Result 2: In the judgment game, $\partial r/\partial r'' = \partial p/\partial p''$: the causal effect of second-order expectations on behavior is equal in reward and punishment choices.

In model 1 in table 1, we do a Wald test of equality on the causal effects of second-order expectations on reward and punishment choices. The p-value is 87%.

Result 3a: Result 1 differs substantially if we use data from the elicitation treatment, suggesting that elicited second-order expectations are endogenous. Result 2 does not differ.

Model 2 in table 1 uses data from the elicitation treatments. The estimated causal effect of second-order expectations on choices is positive for reward and punishment choices (which mimics Dufwenberg and Gneezy (2000)). It is both economically and statistically significant. The Spearman rank correlation is 0.57 (p-value < 1%) for reward choices and 0.61 (p-value < 1%) for punishment choices. Note that result 2 is unaffected: a Wald test of equality on the causal effects of second-order expectations on reward and punishment choices yields a p-value of 27%.

Result 3b: Results 1 and 2 are not affected by using data from the inducement treatment rather than the instrumental treatment.[20]

Model 3 in table 1 uses data where we report directly the first-order expectations of senders to responders without notifying senders of our intent to do so, and notifying responders of this fact. The estimated causal effect of second-order expectations on reward and punishment choices is positive and statistically insignificant. The Spearman rank correlation is 0.24 (p-value = 22%) for reward choices and 0.08 (p-value = 66%) for punishment choices. A Wald test of equality on the causal effects of second-order expectations on reward and punishment choices yields a p-value of 79%. Thus results 1 and 2 are unaffected.

4. Discussion

Our elicitation results for reward choices are consistent with Dufwenberg and Gneezy (2000) and the subsequent studies (fn 3). These studies interpreted the observed behavioral confirmation as support

[19] They used the observer first-order expectations to aid the responders in formulating accurate second-order expectations.

[20] This suggests that there is no implicit deception effect.

for the guilt-aversion model.[21] Charness and Dufwenberg (2006) acknowledged the potential endogeneity of second-order expectations, which led Ellingsen et al. (2009) to randomly induce second-order expectations. Ellingsen et al. (2009) found an insignificant effect of second-order expectations and concluded that: "… guilt aversion is accordingly smaller than previously thought," (p15).

We find similar results when using the inducement method or when instrumenting. However our interpretation differs. The guilt-aversion model is not the only model linking second-order expectations to actions.[22] Several models explain reward and punishment in terms of a desire to reciprocate intentions, and second-order expectations play an important role in the responder assessing the sender's intentions (see fn 5; also for rewards standards, see Gneezy and Guth (2003)).

To understand why, let us reconsider the lost wallet game in figure 1 in light of the intentional reciprocity model. If the responder thinks that the sender sent over the \$10 expecting nothing back $(y'' = 0)$, then the responder will regard the sender's act as being kind, and therefore deserving of reward. On the other hand, if the responder thinks that the sender was expecting everything back $(y'' = 30)$, then the responder will regard the sender as selfish and undeserving of reward. In the case of reward choices, the desire to reciprocate intentions implies behavioral disconfirmation.

Thus in the case of reward choices, intentional reciprocity and guilt-aversion make opposing predictions about the sign of the causal effect of second-order expectations on choices (the studies in fn 3 note this). However we should not interpret the sign of the estimated causal effect as implying support for one and – more importantly – evidence against the other. Rather, for two reasons, we should allow ourselves to interpret the estimated sign as describing which of the two effects *is stronger* in the reward decision being studied.

First, the cognitive processes underlying the two models of behavior are not mutually exclusive. It is perfectly plausible to think of a responder balancing the desire to reciprocate selfish intentions by withholding a reward with a desire to avoid letting the sender down by failing to reward.

Second, there is a large body of evidence in favor of both models (see the cited papers). The models generate testable predictions far beyond the narrow confines of the causal effect of second-order expectations in trust games. The conclusion is that humans are clearly both guilt-averse and that they have a propensity to reciprocate intentions. The question is therefore: which of these effects is stronger in trust games?

Combining our results with those in Ellingsen et al. (2009) and those generated by applying instrumental methods to Bacharach et al.'s (2007) data, our answer is that the two effects seem to cancel each other out (though this is at odds with Reuben et al.'s (2008) results). The data do not permit us to infer the absolute strength of either guilt-aversion or intentional reciprocity – only their relative strengths.

[21] For more on guilt and guilt-aversion, see Baumeister et al. (1994), Battigalli and Dufwenberg (2007) and Vanberg (2008).

[22] Guilt-aversion is not the only model of behavioral confirmation. See Jussim (1986), Pelletier and Vallerand (1996) and Heilman and Alcott (2001).

In the case of punishment choices, the theoretical picture is a little more blurred. In the guilt-aversion model, a responder first considers what she thinks that the sender expects the sender's payoff to be, $\pi_s^{''}$. Giving the sender less than she thinks that the sender is expecting generates disappointment $d = \max(0, \pi_s^{''} - \pi_s)$.

A responder deviates from her materialistic optimum to avoid generating positive disappointment. In reward choices, the materialistic optimum (zero reward) minimizes the sender's payoff and so guilt-aversion may imply some reward (behavioral confirmation). In contrast, in punishment choices, the materialistic optimum (zero punishment) maximizes the sender's payoff and so guilt-aversion is consistent with zero punishment: increasing punishment can only increase disappointment. Thus punishment decisions are associated with neither behavioral confirmation nor behavioral disconfirmation.

The situation becomes more complicated if the responder has another reason for imposing punishment (e.g., wanting to equalize payoff outcomes; see Fehr and Schmidt (1999) and Bolton and Ockenfels (2000)). In this case, guilt-aversion makes the responder want to punish the sender less than she thinks that the sender expects to be punished. The second-order expectation of punishment acts as an upper bound on punishment, which is a form of behavioral confirmation.

Models of intentional reciprocity are even more ambiguous in the punishment domain. If the sender plays $unkind$, does thinking that the sender expects to be punished more make her more or less deserving of punishment? Intuitively the answer is not clear, and the predictions of the theoretical models in the economic literature reflect this (see Appendix 2 for derivations of what follows).

In the Dufwenberg and Kirchsteiger (2004) model, the responder's focus is her payoff compared to what else she could have earned. Accordingly, the more the responder thinks that the sender expects to get punished, the lower the responder thinks that the sender expected the responder's payoff to be. In words, the responder thinks: "I can't believe you played $unkind$ while expecting me to punish you so much that my payoff will be lower than it could have been; just for that I'll punish you a lot!" The result is behavioral confirmation in punishment.

In the Falk and Fischbacher (2006) model, the responder uses her payoff vs. that of the sender as the benchmark for evaluating intentions. Thus depending on how expensive punishment is, the responder can think one of two things: (1) "I can't believe you played $unkind$ while expecting me to punish you so much that my payoff will be a lot lower than yours; just for that I'll punish you a lot!," (behavioral confirmation) or (2) "You played $unkind$ while expecting me to punish you so much that my payoff would be only a little lower than yours; just for that I won't punish you very much," (behavioral disconfirmation).

The key is whether punishment brings the players' payoffs closer to each other or further away. In the judgment game when the marginal cost of punishment is low (which occurs at low punishment levels due to the quadratic cost), we have behavioral disconfirmation. At high punishment levels, the marginal cost is high and so we have behavioral confirmation. Yet even if one were to use a game with a constant marginal cost of punishment to ensure an unambiguous prediction, this would still ultimately be ad hoc.

The model will always be sensitive to arbitrary factors such as whether the benchmark is the absolute or proportionate payoff inequality, or whether there is non-linear distaste for payoff inequality.

The bottom line is that models of intentional reciprocity, and to some extent models of guilt-aversion, do not make unambiguous predictions concerning the causal effect of second-order expectations on punishment decisions. Using instrumental variables, we found an insignificant effect of second-order expectations on punishment decisions. In light of the ambiguity of the models, it is difficult to interpret this result. We do not regard this as a shortcoming of the design; recall that the study is principally motivated by a desire to measure the causal effect of second order expectations on reward and punishment decisions.[23]

Finally, we turn our attention to the apparent endogeneity of elicited second-order expectations. Why might second-order expectations be correlated with reward and punishment choices? We subscribe to Charness and Dufwenberg's (2006) appeal to the false consensus effect: people project their way of thinking on to others. If you want to reward or punish someone for whatever reason, you (falsely) believe that others think the same way as you, and eliciting your beliefs will expose this.

5. Summary and conclusion

Following Dufwenberg and Gneezy (2000), this paper addressed three questions:

1. When deciding on whether to reward or punish someone, how does how you think others expect you to behave affect your decision?
2. Does it depend upon whether others expect you to reward them vs. punish them?
3. What is the interpretation of such a causal effect?

The manipulability of second-order expectations and the importance of reward and punishment decisions mean that these are important questions. One should note that we attempted to answer these only within the context of the judgment game in a laboratory setting (see Gneezy and List (2006) for a field reciprocity game), which is one of many variants of the trust game. Nevertheless we were reassured by the consistency of our results with previous studies (or data garnered from previous studies).

We found that second-order expectations had a (statistically) zero causal effect on both reward and punishment decisions. We also found that it is important to instrument for second-order expectations because they are endogenous.

From the perspective of optimal incentive schemes, this suggests that there is no gain to exogenously manipulating second-order expectations: in one-shot environments, changing how others think you are

[23] As mentioned earlier in the discussion, theories of guilt-aversion and intentional reciprocity produce a range of testable hypotheses that extend far beyond the relationship between second-order expectations and actions. See the social psychology studies that we cite for tests that do not encounter these interpretation issues.

expecting to get rewarded or punished has no impact on the incidence of either. This is consistent with Dufwenberg and Charness' (2006) finding that messages sent by first-movers (senders) did not affect second-mover (responder) behavior. Since such manipulation can be costly, resources devoted to improving incentive schemes are best directed elsewhere.

In light of the substantial evidence on the importance of both guilt-aversion and intentional reciprocity to reward decisions (much of which is unrelated to second-order expectations), we interpreted our results as implying mutual cancellation of the two mechanisms. One need not treat the two explanations as mutually exclusive. In punishment decisions, generating testable predictions from the models requires ad hoc assumptions, and so explaining the data is harder. Ultimately, since punishment is costly to the punisher, there is no intuitively obvious answer to the question: "If I think that you are expecting me to punish you, does that make me more or less likely to punish you?;" this remains an interesting avenue for future research.

Our contributions have been extending the analysis of the causal effect of second-order expectations on behavior to the domain of punishment choices, and, in the context of reward behavior, asserting the mutual inclusivity of guilt-aversion and intentional reciprocity in explaining the data. Finally, to the best of our knowledge, we are the first to use instrumental variable methods to ensure exogenous variation in second-order expectations.

References

Ames, D., F. Flynn and E. Weber (2004). "It's the thought that counts: on perceiving how helpers decide to lend a hand," Personality and Social Psychology Bulletin. 30, p461-474.

Al-Ubaydli, O. and M. Lee (2009). "An experimental study of asymmetric reciprocity," Journal of Economic Behavior and Organization. Forthcoming.

Bacharach, M., G. Guerra and D. Zizzo (2007). "The self-fulfilling property of trust: an experimental study," Theory and Decision. 63, p349-388.

Battigalli, P. and M. Dufwenberg (2007). "Guilt in games," American Economic Review. 97, p170-176.

Baumeister, R., A. Stillwell and T. Heatherton (1994). "Guilt: an interpersonal approach," Psychological Bulletin. 115, p243-267.

Berg, J., J. Dickhaut and K. McCabe (1995). "Trust, reciprocity and social history," Games and Economic Behavior. 10, p122-142.

Blount, S. (1995). "When social outcomes aren't fair: the effect of causal attributions on preferences," Organizational Behavior and Human Decision Processes. 63, p131-144.

Bolton, G. and A. Ockenfels (2000). “ERC: a theory of equity, reciprocity and competition,” American Economic Review. 90, p166-193.

Charness, G. and M. Dufwenberg (2006). “Promises and partnership,” Econometrica. 74, p1579-1601.

Cox, J., D. Friedman and S. Gjerstad (2007). “A tractable model of reciprocity and fairness,” Games and Economic Behavior. 59, p17-45.

Dufwenberg, M. and U. Gneezy (2000). “Measuring beliefs in an experimental lost wallet game,” Games and Economic Behavior. 30, 163-182.

Dufwenberg, M. and G. Kirchsteiger (2004). “A theory of sequential reciprocity,” Games and Economic Behavior. 47, p268-298.

Ellingsen, T., M. Johannesson, S. Tjotta and G. Torsvik (2009). “Testing guilt aversion,” Games and Economic Behavior. Forthcoming.

Falk, A. and U. Fischbacher (2006). “A theory of reciprocity,” Games and Economic Behavior. 54, p293-315.

Fehr, E. and K. Schmidt (1999). “A theory of fairness, competition and cooperation,” Quarterly Journal of Economics. 114, p817-868.

Fershtman, C. and U. Gneezy (2001). “Discrimination in a segmented society: an experimental approach,” Quarterly Journal of Economics. 116, p351-377.

Geanakoplos, J., D. Pearce and E. Stacchetti (1989). “Psychological games,” Games and Economic Behavior. 1, p60-79.

Gneezy, U. (2005). “Deception: the role of consequences,” American Economic Review. 95, p384-394.

Gneezy, U. and W. Guth (2003). “On competing rewards standards – an experimental study of ultimatum bargaining,” Journal of Socio-Economics. 31, p599-607.

Gneezy, U. and J. List (2006). “Putting behavioral economics to work: testing for gift exchange in labor markets using field experiments,” Econometrica. 74, p1365-1384.

Greenberg, M. and D. Frisch (1972). “Effect of intentionality on willingness to reciprocate a favor,” Journal of Experimental Social Psychology. 8, p99-111.

Guerra, G. And D. Zizzo (2004). “Trust responsiveness and beliefs,” Journal of Economic Behavior and Organization. 55, p25-30.

Heilman, M. and V. Alcott (2001). “What I think you think of me: women’s reactions to being viewed as beneficiaries of preferential selection,” Journal of Applied Psychology. 86, p574-582.

Holm, H. and P. Nystedt (2005). “Intra-generational trust – a semi-experimental study of trust among different generations,” Journal of Economic Behavior and Organization. 58, p403-419.

Jussim, L. (1986). “Self-fulfilling prophecies: a theoretical and integrative review,” Psychological Review. 93, p429-445.

McCabe, K., M. Rigdon and V. Smith (2003). “Positive reciprocity and intentions in trust games,” Journal of Economic Behavior and Organization. 52, p267-275.

McCullough, M., R. Emmons, S. Kilpatrick and D. Larson (2001). “Is gratitude a moral effect?,” Psychological Bulletin. 127, p249-266.

Offerman, T. (2002). “Hurting hurts more than helping helps,” European Economic Review. 46, p1423-1437.

Offerman, T., J. Sonnemans and A. Schram (1996). “Value orientations, expectations and voluntary contributions in public goods,” Economic Journal. 106, p817-845.

Pelletier, L., and R. Vallerand (1996). “Supervisors’ beliefs and subordinates’ intrinsic motivation: a behavioral confirmation analysis,” Journal of Personality and Social Psychology. 71, p331-340.

Pruitt, D. (1968). “Reciprocity and credit building in a laboratory dyad,” Journal of Personality and Social Psychology. 8, p143-147.

Rabin, M. (1993). “Incorporating fairness into game theory and economics,” American Economic Review. 83, p1281-1302.

Reuben, E., P. Sapienza and L. Zingales (2008). “Is mistrust self-fulfilling?,” Working paper, Northwestern University.

Schopler, J. and V. Thompson (1968). “Role of attribution processes in mediating amount of reciprocity for a favor,” Journal of Personality and Social Psychology. 10, p243-250.

Tesser, A. R. Gatewood and M. Driver (1968). “Some determinants of gratitude,” Journal of Personality and Social Psychology. 9, p233-6.

Vanberg, C. (2008). “Why do people keep their promises? An experimental test of two explanations,” Econometrica. 76, p1467-1480.

Appendix 1: Results from Bacharach et al. (2007)

		G game		K game		N game	
		Responder		Responder		Responder	
		y = 1	y = 0	y = 1	y = 0	y = 1	y = 0
Sender	x = 1	3,3	-3,4.5	3,3	-3,4.5	3,3	-3,4.5
	x = 0	0,3	0,3	0,0	0,0	-1.5,0	-1.5,0

Figure A1: Trust games in Bacharach et al. (2007)

The first payoff in each pair is the sender's payoff. We have renamed strategies and players to facilitate comparison with the other games presented in this paper.

Bacharach et al. (2007) use three variants of the trust game, show in figure A1. Note that only the option reward exists; there is no punishment.

	Model 1	**Model 2**
Estimation method	Probit	IV probit
Constant	0.88***	-0.55
Standard error	(0.34)	(0.49)
G game dummy	-0.26	0.23
Standard error	(0.45)	(1.12)
K game dummy	0.09	0.36
Standard error	(0.48)	(0.69)
y'' x G game dummy	0.01	-0.02
Standard error	(0.01)	(0.03)
y'' x K game dummy = [A]	0.02***	-0.01
Standard error	(0.01)	(0.01)
y'' x N game dummy = [B]	0.02***	-0.01
Standard error	(0.01)	(0.01)
Observations	160	160
Pseudo R^2	0.10	-

Table A1: Probit results from Bacharach et al. (2007)

The dependent variable in all models is a dummy variable taking the value 1 if the responder plays 1. 'G game dummy' and 'K game dummy' are dummies denoting the game. y'' denotes the responder's second-order expectation of y. Asterices denote statistical significance (* = 10%, ** = 5%, *** = 1%).

As a slight abuse of notation, let y'' denote the responder's second-order expectation of the probability that she will play $y = 1$.

Table A1 shows the probit results. Model 1 is basically table IV from Bacharach et al. (2007) without controls for suspicious observations or demographics. The estimated causal effect of second-order expectations is positive in all three games, and it is significant for the K and N games.

In model 2, we use the average guess of neutral observers as an instrument for elicited second-order expectations. The estimate causal effect becomes negative and statistically insignificant in all three games.

Appendix 2: Economic models of intentional reciprocity

Let the sender be player 1 and the responder be player 2. In the judgment game, Dufwenberg and Kirchsteiger (2004) preferences for the responder take the form:

$$u_2 = \pi_2 + \rho^{DK}\pi_1 k^{DK}(r'', p'')$$

Falk and Fischbacher (2006) preferences for the responder take the form:

$$u_2 = \pi_2 + \rho^{FF}\pi_1 k^{FF}(r'', p'')$$

π_i denotes the payoff of player i. k denotes the responder's assessment of how kind the sender is. $\rho \geq 0$ denotes the reciprocity parameter. When $k < 0$, the responder is willing to punish at the margin. When $k > 0$ the responder is willing to reward at the margin. Reward/punishment choices are increasing in $|k|$. We refer the reader to the two papers for the background on these preferences.

Fact: When the sender plays $kind$, in both models, r is everywhere decreasing in r''.

Proof: The kindness terms are:

$$k^{DK} = \left[24 - \frac{7}{100}r''^{2}\right] - \left[\frac{1}{2}\left(\left(24 - \frac{7}{100}r''^{2}\right) + \left(12 - \frac{7}{100}p''^{2}\right)\right)\right]$$
$$= \left[\frac{1}{2}\left(\left(24 - \frac{7}{100}r''^{2}\right) - \left(12 - \frac{7}{100}p''^{2}\right)\right)\right]$$

$$k^{FF} = \left[24 - \frac{7}{100}r''^{2}\right] - [10 + r''] = 14 - 2r'' - \frac{7}{100}r''^{2}$$

Both of which are clearly decreasing in r''. ■

Fact: When the sender plays $kind$, in the Dufwenberg and Kirchsteiger (2004) model, p is everywhere increasing in p'', while in the Falk and Fischbacher (2006) model, p is non-monotonic in p''.

Proof: The kindness terms are:

$$k^{DK} = \left[12 - \frac{7}{100}p''^{2}\right] - \left[\frac{1}{2}\left(\left(24 - \frac{7}{100}r''^{2}\right) + \left(12 - \frac{7}{100}p''^{2}\right)\right)\right]$$
$$= \left[\frac{1}{2}\left(\left(12 - \frac{7}{100}p''^{2}\right) - \left(24 - \frac{7}{100}r''^{2}\right)\right)\right] < 0$$

$$k^{FF} = \left[12 - \frac{7}{100}p''^{2}\right] - [16 - p''] = -4 + p'' - \frac{7}{100}p''^{2} < 0$$

Clearly $|k^{DK}|$ is everywhere increasing in p''. $|k^{FF}|$ is decreasing in p'' for $p'' \in [0, 50/7]$ and increasing for $p'' \in [50/7, 12]$. ■

Appendix 3: Experimental instructions

A. Main instructions

Thank you for agreeing to participate in today's experiment. You will be undertaking three different decision-making tasks. In each of the three tasks, there are two subject types: BLUES and REDS. You will either be a BLUE for all three tasks or a RED for all three tasks. Before entering the room, you have drawn a chip from the bag that determined which type you will be. REDS are in this room and BLUES will be in a different room.

In each task, a BLUE will be paired with a RED. Your partner will be randomly selected and you will never know each other's identity. After each task, you will be randomly reassigned a new partner. You will never have the same partner for more than one task. Each task will be undertaken only once.

In each task, the decisions that you and others make will affect your earnings. Depending on your decisions, you may then earn a considerable amount of money. However, only one of the three tasks will actually be used to determine your earnings. At the end of the three tasks, your earnings will be paid to you individually and in private. In each room, we will ask one of the participants to draw a card to determine which of the three tasks will be used to calculate your earnings.

Earnings in each task are denominated in points. Every 3 points are worth $1. So, for example, if in the task that ends up being used to calculate your earnings you earn 15 points then that will correspond to $5. Including the time it takes to calculate and deliver earnings, this session should last about 1 hour. Are there any questions?

From now until the end of the session, unauthorized communication of any nature with other participants is prohibited. At any time, if you have questions, please raise your hand and the monitor will come to you.

Welcome to the triangle decision task. Recall that each of you has been anonymously paired with a randomly selected member of the BLUE room. In this decision task, you will make a choice and your partner will make a choice. You choose in response to your partner's choice.

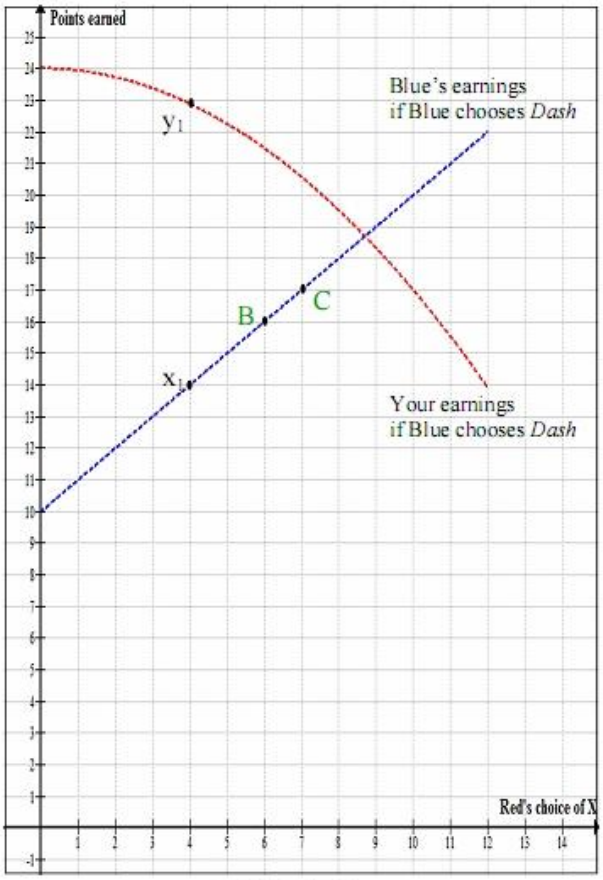

FIGURE 1
If BLUE chooses *Dash*

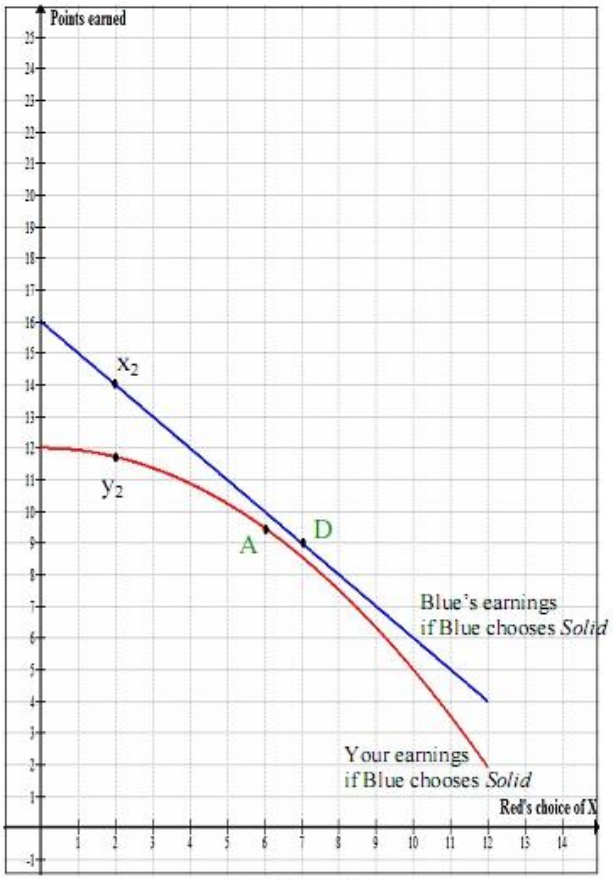

FIGURE 2
If BLUE chooses *Solid*

Your BLUE partner's decision:

- Your partner chooses first between *Dash* and *Solid*.
- If he/she chooses *Dash*, your earnings increase by *24* points and your partner's earnings increase by *10* points.
- If he/she chooses *Solid*, your earnings increase by *12* points and your partner's earnings increase by *16* points.
- You get to make your choice in response to your BLUE partner's choice.

Your decision:

- If your BLUE partner plays *Dash*, you can **increase** his/her earnings by any amount in the range 0 to 12 points
- If your BLUE partner play *Solid*, you can **decrease** his/her earnings by any amount in the range 0 to 12 points
- Changing your earnings comes at a cost to you: to change his/her earnings by *X* costs you *7%* of X^2.

The supplementary diagrams represent these earnings to you. The red curves are your final earnings and the blue lines are your partner's final earnings. The dashed line corresponds to your partner picking *Dash* (figure 1) and the solid line to your partner picking *Solid* (figure 2). How right the earnings are depends upon your choice of *X*, which is the horizontal axis.

For example, if your partner selects *Dash* and then you select *X = 4*:

- Your final earnings are *22.9* points (*24 – 7%* of 4^2, point y_1 in figure 1)
- Your partner's earnings are *14* points (*10 + 4,* point x_1 in figure 1)
- It costs you *1.1* to increase their earnings by *4*.

For example, if your partner selects *Solid* and then you select *X* = 2:

- Your final earnings are *11.7* points (*12 – 7%* of $(-2)^2$, point y_2 in figure 2)
- Your partner's earnings are *14* points (*16 – 2*, point x_2 in figure 2)
- It costs you *0.3* to decrease their earnings by *2*.

To make sure that you understand how to calculate the final earnings, please answer the following 2 questions (note that you DO NOT need any mathematics to answer these questions - you are simply locating points on the figures):

1. Assume a BLUE selected *Dash* and, in response, the RED partner selects *X = 7*. Which point (among *A*, *B*, *C*, *D* in either figure 1 or 2) represents BLUE's final earnings? (Circle your answer)

a) *A* b) *B* c) *C* d) *D*

2. Assume a BLUE selected *Solid* and, in response, the RED partner selects *X = 6*. Which point (among *A*, *B*, *C*, *D* in either figure 1 or 2) represents RED's final earnings? (Circle your answer)

a) *A* b) *B* c) *C* d) *D*

Rather than seeing your BLUE partner's actual choice and then you making your choice, we will ask you for two pieces of information: your choice if it turns out that your partner picks *Dash*, and your choice if it turns out that your partner picks *Solid*. At the end of the experiment, we will tell you which choice your partner actually made. Out of the two choices you make, the one we use to calculate your earnings will be the one that corresponds to your partner's actual choice.

Soon you will make your choices. Once you have made your choices, we will use your decisions to calculate 4 different average earnings:

Suppose ALL the BLUES decided to select *Dash*:

1. What will be the BLUES' average earnings?
2. What will be the REDS' average earnings?

Both of these depend on what the REDS say they would do if their partners select *Dash.*

Suppose ALL the BLUES decided to select *Solid*:

3. What will be the BLUES' average earnings?
4. What will be the REDS' average earnings?

Both of these depend on what the REDS say they would do if their partners select *Solid*

In previous sessions, after they decided between *A* and *B*, we asked the BLUES to predict what these 4 average earnings would be, telling them that they will be rewarded for their accuracy. Those earnings were extra earnings for that task.

We will now show each of you the average of what 10 BLUES from previous sessions predicted. It is important to note that these predictions were never shown to the actual partners of the BLUES. Like all GMU experiments, there is no deception in our experiment.

On the choice card, please record your choice when your partner selects *Solid* and your choice when your partner selects *Dash*.

After they decided between *Dash* and *Solid*, we asked the BLUES in the other room to predict what the above 4 average earnings would be, telling them that they will be rewarded for their accuracy. These earnings will be extra earnings for this task.

Now we want you to predict the average of the BLUES' guesses of these 4 average earnings. In other words, what do you think that the BLUES predicted that the REDS in this room would do? We will also reward you for your accuracy. These will be extra earnings for this task. The earnings will be calculated on the basis of a formula. It is not important that you have mathematical insight into this formula. But it is important that you realize that your average earnings will be maximized if you report your expectations truthfully. It is to your advantage to report your expectations honestly. You will always earn a positive amount for your guess, but the more accurate your guess the more you will earn.

For completeness, the formula will be given in a handout.

It is now time to make your predictions. On the prediction card, please record your predictions for these averages.

B. Predictions supplement

These are the 4 average earnings:

- Suppose ALL the BLUES decided to select *Dash*. What will be the BLUES' average earnings? This average depends on what the REDS say they would do if their partners select *Dash*.

- The average of the BLUES' guesses is x_{Dash}^{BLUE} and your prediction of the average of the BLUES' guesses is $\hat{x}_{Dash}^{BLUE}$.
- Again suppose ALL the BLUES decided to select *Dash*. What will be the REDS' average earnings?
- The average of the BLUES' guesses is x_{Dash}^{RED} and your prediction of the average of the BLUES' guesses is $\hat{x}_{Dash}^{RED}$.
- Now instead, suppose ALL the BLUES decided to select *Solid*. What will be the BLUES' average earnings? This average depends on what the REDS say they would do if their partners select *Solid*.
- The average of the BLUES' guesses is x_{Solid}^{BLUE} and your prediction of the average of the BLUES' guesses is $\hat{x}_{Solid}^{BLUE}$.
- Finally suppose ALL the BLUES decided to select *Solid*. What will be the REDS' average earnings?
- The average of the BLUES' guesses is x_{Solid}^{RED} and your prediction of the average of the BLUES' guesses is $\hat{x}_{Solid}^{RED}$.

The number of points that you will earn is:

$$8 - 0.1 \times \left(x_{Dash}^{BLUE} - \hat{x}_{Dash}^{BLUE}\right)^2 - 0.1 \times \left(x_{Dash}^{RED} - \hat{x}_{Dash}^{RED}\right)^2 - 0.1 \times \left(x_{Solid}^{BLUE} - \hat{x}_{Solid}^{BLUE}\right)^2 - 0.1 \times \left(x_{Solid}^{RED} - \hat{x}_{Solid}^{RED}\right)^2$$

If this figure is less than zero, we will reset it to zero so that you cannot lose points by guessing badly – you can only gain points.

C. Changes in rubric for inducement treatment

In section bounded by horizontal lines, insert the following as a replacement:

After they decided between *Dash* and *Solid*, we asked the BLUES to predict what these 4 average earnings would be, telling them that they will be rewarded for their accuracy. These earnings will be extra earnings for this task.

We will now show each of you what your individual BLUE partner predicted. It is important to note that the BLUES WERE NOT TOLD that you would see their predictions.