Exploiting the Guilt Aversion of Others - Do Agents do it and is it Effective?[*]

Eric Cardella[†]

July 17, 2015

Abstract

The general idea of guilt aversion is that agents may be motivated to avoid letting others down, even at the expense of their own material payoff. Several experimental studies have documented behavior that is consistent with agents exhibiting guilt averse motivations in social interactions. However, there are strategic implications of guilt aversion, which can impact economic outcomes in important ways, that have yet to be explored. I introduce a game that admits the possibility for agents to induce guilt upon others in a manner consistent with the method posited by Baumeister, Stillwell, and Heatherton (1994). This game enables me to experimentally test whether agents attempt to exploit the guilt aversion of others by inducing guilt upon them, and whether agents are actually susceptible to this exploitation. Additionally, the design enables me to test whether agents exhibit higher degrees of trust when they are given such an opportunity to exploit the guilt aversion of others. The data suggests that agents do not attempt to fully exploit the guilt aversion of other agents by inducing guilt upon them; however, the data suggests that agents *would have been* susceptible to guilt induction.

Keywords: guilt aversion, trust, psychological game theory, experiment

JEL Codes: C72, C91, D03, D80

[*] I would like to thank my dissertation advisor Martin Dufwenberg for helpful comments. I would also like to thank Pierpaolo Battigalli, Anna Breman, Gary Charness, Price Fishback, Paul Heidhues, Tamar Kugler, John Wooders, participants at the ESA conference in Tucson, participants from BABEEW at Santa Clara University, and seminar participants at the University of Arizona, Rochester Institute of Technology, University of Mississippi and Bowdoin College for additional helpful comments. This project was funded by the Economic Science Laboratory (ESL) and the National Science Foundation (NSF) (grant # SES-0819780).

[†] Department of Energy, Economics, and Law; Rawls College of Business, Texas Tech University, Lubbock, TX 79409, Telephone: (806) 834-7482; Email: eric.cardella@ttu.edu.

1 Introduction

The results from a growing body of experimental literature suggest that economic agents may not be solely motivated to maximize their own material payoffs. One example of such a "non-selfish" behavioral motivation is guilt aversion. The general idea of guilt aversion is that an agent would suffer disutility, in the form of guilt, from hurting or letting down another agent, relative to that other agent's expectations;[1] thus, a guilt averse agent may be motivated to avoid hurting or letting down that other agent, even at the expense of his/her own material payoff, to assuage the guilt feeling. Behavior consistent with agents exhibiting guilt aversion has been documented in several experimental studies. For example, Dufwenberg and Gneezy (2000), Charness and Dufwenberg (2006, 2010), Bacharach et al. (2007), Reuben et al. (2009); and Attanasi et al. (2014b) find evidence of guilt aversion using variations of two-player experimental "trust" games (Berg et al., 1995).[2] Specifically, these papers show that the amount of money the trustee (second-mover) returns to the trustor (first-mover) is positively correlated with how much the trustor expects to get back (or how much the trustee thinks the trustor expects back). The idea being that guilt averse trustees give back more (forgo their own material payoff) the more their trustor expects back in order to avoid the guilt that would result from failing to meet the trustor's expectations.

In light of the experimental evidence of exhibited guilt aversion, it is important to consider the richer set of interpersonal strategic implications that can arise when agents are motivated by guilt aversion. In particular, the guilt aversion of one agent can influence the behavior of other agents in important ways. In certain social interactions, the possibility may arise for agents to behave opportunistically and exploit the guilt aversion of others. Charness and Dufwenberg (2006), in their concluding remarks, point to such a possibility by raising the question, "do people manipulate the guilt aversion of others in self-serving ways?" (p. 1595) Given the opportunity and incentive, agents could attempt to influence the behavior of guilt averse others by strategically inducing guilt upon these agents, i.e., increasing the amount of guilt they would feel. Consequently, guilt averse agents may be more motivated to respond in kind to avoid hurting or letting down those agents that had induced guilt upon them. These interpersonal implications of guilt aversion can impact strategic decision making and, consequently, economic outcomes in important ways that have yet to be explored.

The goal of this study is to explore some of these interpersonal implications of guilt aversion. Specifically, this paper introduces an experimental design aimed at shedding light on the following three questions: First, do economic agents attempt to exploit the guilt aversion of other agents in self-serving ways by strategically inducing guilt upon those other agents? Second, is strategic guilt induction effective at influencing the behavior of other agents, i.e., are agents susceptible to this exploitation? Third, does having the opportunity to induce guilt upon other agents impact strategic behavior?

Although previously unexplored in the economics literature, the interpersonal implications of guilt aversion have been studied and documented in the psychology literature (Vangelisti et al., 1991; Baumeister et al. (BSH henceforth), 1994, 1995; Tangney and Fischer, 1995; de Hooge et al.,

[1] For a more thorough discussion of the psychological foundations of guilt, I refer interested readers to Hoffman (1982), Tangney (1990), Tangney and Fischer (1995), Baumeister et al. (1994), (1995), and Tangney and Dearing (2002). Formal game theoretic models of guilt aversion have been developed by Dufwenberg (2002), for a specific game, and later by Battigalli and Dufwenberg (2007) for a general class of games.

[2] Other studies documenting guilt averse behavior include Nelissen et al. (2011) in an ultimatum game, and Dufwenberg et al. (2011) in a public goods game. Relatedly, Bellemare et al. (2011) document direct evidence of agents willing to pay (postive monetary amounts) to avoid letting others down. On the contrary, Ellingsen et al. (2010) find little experimental support for guilt aversion.

2011). In particular, BSH (1994) argue that one of the primary functions of guilt is to motivate others to behave in a more desirable way. In their study, BSH (1994) note that "we observed ample evidence of the hypothesized function of guilt as an interpersonal influence technique: People induced guilt to get another person to comply with their wishes." (p. 249) Similarly, Vangelisti et al. (1991) argue that people induce guilt "primarily to achieve their own end-to persuade their listeners to do or not to do something." (p. 33) These psychology studies provide foundational insights regarding the interpersonal functions of guilt in social relationships by drawing conclusions from non-incentivized personal narratives and surveys. However, these functions of guilt may not be restricted to social interactions; guilt may also function as an "interpersonal influence technique" in strategic economic interactions. An incentivized experimental game provides a suitable platform for investigating these interpersonal implications of guilt aversion in economic settings.

Before I proceed, I pause to highlight some of the economic settings where strategic guilt induction could be relevant in terms of influencing the behavior of others and, consequently, impact outcomes. In contracting environments, guilt induction may allow a disadvantaged party to influence the behavior of an advantaged counter-party. For instance, a contracted firm that had made relationship-specific investments could possibly thwart opportunist re-contracting and hold-up by conveying to the counter-party firm the loss in profits associated with such a hold-up. In the workplace, managers could induce guilt upon employees to mitigate shirking by conveying to employees how their sub-standard effort adversely affects other employees.[3] In academia, assistant professors could *possibly* induce guilt upon journal editors to get a more timely review decision on a submitted paper by *gently* informing the editor, at the time of submission, that his/her tenure review is rapidly approaching and a lengthy review period could hinder his/her tenure prospects.[4] Many economic settings, like these mentioned, permit the possibility to induce guilt upon others. Hence, a deeper understanding of the strategic interpersonal implications of guilt aversion is required to better ascertain how the guilt aversion of agents will impact outcomes in such settings; the insights gleaned from this paper are intended to help with this understanding.

In order to investigate whether agents strategically induce guilt upon others and its potential effectiveness, it is crucial to first identify how agents can attempt to induce guilt upon others. For this, I draw foundational insights from BSH (1994), who posit the following method for how people induce guilt in others: "If Person A wants Person B to do something, A may induce [more] guilt in B by conveying how A suffers [how much A is let down] over B's failure to act in the desired fashion" (p. 247). In regards to strategic economic settings, this proposed method by BSH (1994) would correspond to Person A conveying how low his/her payoff would be, i.e., how he/she suffers, as a result of Player B choosing an undesirable action toward Player A.[5] Note that this method for inducing guilt implicitly requires that (i) Person A has private information about his/her own payoff and the degree to which he/she may suffer from Player B's action, and (ii) Person A has the possibility to convey such private information to Person B.[6] Previous studies that have investigated

[3] Sub-standard effort by employees is likely to result in lower profits for a firm. Assuming that bonuses are increasing in firm profits, then lower profits would lead to lower bonuses for all employees. Thus, shirking by one employee could adversely affect the well-being of other employees.

[4] This example was inspired by an editor who revealed to me that some assistant professors do actually inform the editor, at the time of submission, about such an upcoming tenure review!

[5] It is worth noting that this prescribed method for inducing guilt proposed by BSH (1994) is also consistent with implications of the formal guilt model developed by Battigalli and Dufwenberg (2007), as I will discuss in more detail in the subsequent section.

[6] I note here that I am by no means implying that these two conditions are necessary for an agent to *feel* guilt, as is evident by prior studies that have effectively studied guilt in games that have no private information. Rather, my claim is that these two conditions are necessary for one agent to *induce* guilt upon another agent, i.e., for one agent

guilt aversion in strategic settings mostly consider variations of 2-player "trust" games that do not feature either of these properties.[7] Hence, a new game is warranted that provides a rich *enough* strategic structure to allow agents the opportunity to induce guilt upon others.

In this paper, I employ a novel experimental design that uses a 2-player, binary choice trust game featuring both private payoff information and an opportunity to convey this private information. In the game, the privately informed first-mover (Player A) is effectively given an opportunity to convey to the second-mover (Player B) how low his/her payoff would be if Player B fails to act in the desired fashion.[8] This game allows me to then derive testable hypotheses regarding whether Player As attempt to strategically induce guilt upon Player Bs, and whether Player Bs are susceptible to strategic guilt induction. Additionally, I show that an important artifact of this game is the ability to identify behavior consistent with guilt aversion without having to elicit beliefs. The experimental design also includes a second, related trust game that does not feature an opportunity for Player As to induce guilt upon Player Bs. This second game provides a baseline trust measure for Player As, which then allows me to derive a testable hypothesis regarding whether Player As are more trusting of Player Bs when Player As have an opportunity to induce guilt upon Player Bs.

The experimental data seems to suggest that Player Bs are susceptible to the guilt induction of Player As. However, the data reveals little evidence that Player As are attempting to induce guilt upon Player Bs, à la BSH (1994). Interpreted differently, the data reveals that Player As are not attempting to fully exploit the guilt aversion of Player Bs by inducing guilt upon them; although, had they done so, it *would have been* effective at increasing the likelihood of Player B choosing the desired action. Furthermore, I find evidence of marginally significantly higher trust rates by Player As in the trust game where Player As have the opportunity to induce guilt upon Player B, compared to a trust game with no such opportunity.

Furthermore, I show that the derived hypotheses are consistent with predictions of the formal model of "simple" guilt developed by Battigalli and Dufwenberg (2007) (B&D henceforth).[9] I also show that in the game considered, effective guilt induction can be supported in equilibrium under the B&D framework. The paper proceeds with the experimental design and hypothesis development in Section 2. I outline the experimental procedure in Section 3. I present and discuss the results in Section 4, and Section 5 concludes.

to *increase* the amount of guilt another agent may feel from acting in an undesirable way.

[7] A study by Fong et al. (2007) uses a trust game with private information. However, the authors incorporate private information as a means of testing their model of guilt driven reciprocity. Additionally, their trust game features private information for the second mover, while I will consider a trust game with private information for the first mover, as will be shown in the next section. Guth et al. (2013) consider a trust game with uncertain payoffs, but not private information, as a way of testing how payoff uncertainty impacts trust and reciprocity.

[8] I use the term "guilt induction" when referring to this type of strategic attempt by Player A to exploit the guilt aversion of Player B and influence the behavior of Player B. I do this to remain consistent with the psychological foundations and terminology outlined by BSH (1994). However, in relation to the Battigalli and Dufwenberg (2007) model of guilt, it may be more pedagogical to think of this strategic behavior from Player A as "counterfactual" guilt induction. Essentially, Player A is trying to increase the amount of guilt that Player B *would* feel as a result of choosing an action that is undesirable for Player A. This makes the guilt counterfactual in the sense that Player B may never experience the guilt if he/she chooses an action that complies with Player A's desired action.

[9] B&D also model a second form of guilt, "guilt from blame." However, in this paper I will consider only simple guilt and, therefore, for the remainder of the paper when I refer to the guilt model of B&D, I am implicitly referring to the model of simple guilt. The B&D model is an application of the authors' more general theoretical framework developed in Battigalli and Dufwenberg (2009), which extends the psychological game theory framework pioneered by Geanakopolos et al. (1989).

2 Experimental Design

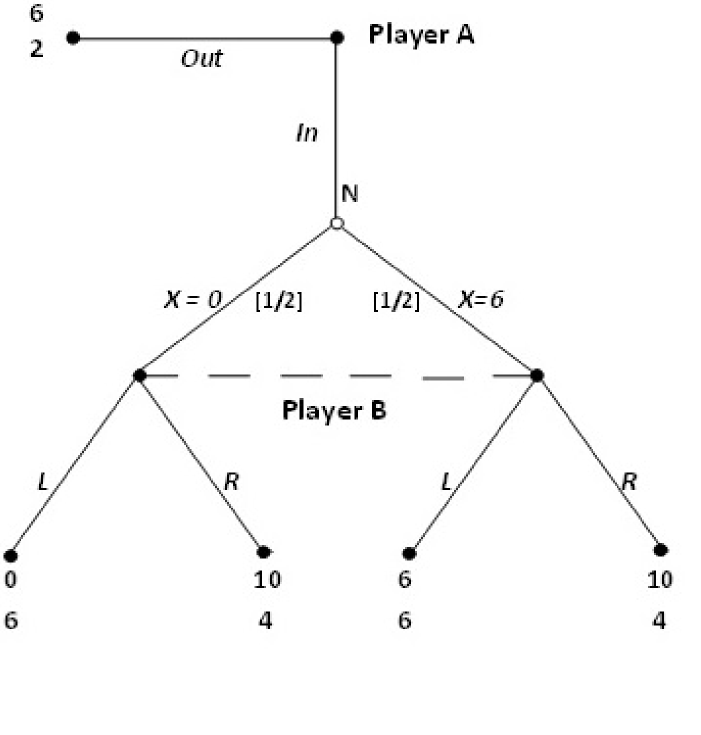
I begin this section by first introducing the two trust games around which the research hypotheses are developed and the experimental design is based. I refer to both games as trust games because they feature a payoff structure indicative of the broader class of trust games.[10] Trust games, in general, allow for the possibility of guilty feelings, which make them a suitable platform for developing and testing the hypotheses of this study relating to the strategic implications of guilt aversion.

2.1 Experimental Trust Games

Uncertain Payoff Trust Game – Γ_{UPT}

Γ_{UPT} is a 2-player, sequential move game. Γ_{UPT} begins with the first mover, Player A, choosing between *In* or *Out.* If Player A chooses *Out,* then the game ends; Player A receives a payoff of 6, and Player B receives a payoff of 2. If Player A chooses *In*, then Player B is called upon to move. Player B must choose between *Left* or *Right.* If Player B chooses *Right,* then the game ends; Player A receives a payoff of 10, and Player B receives a payoff of 4. If Player B chooses *Left,* then the game ends; Player A receives a payoff of X, and Player B receives a payoff of 6. X is a random variable where $prob(X = 0) = 1/2$ and the $prob(X = 6) = 1/2$. At the start of the game, the distribution of X is known to both players. The extensive form of Γ_{UPT} is depicted below in Figure 1.

Figure 1: Extensive Form of Γ_{UPT}



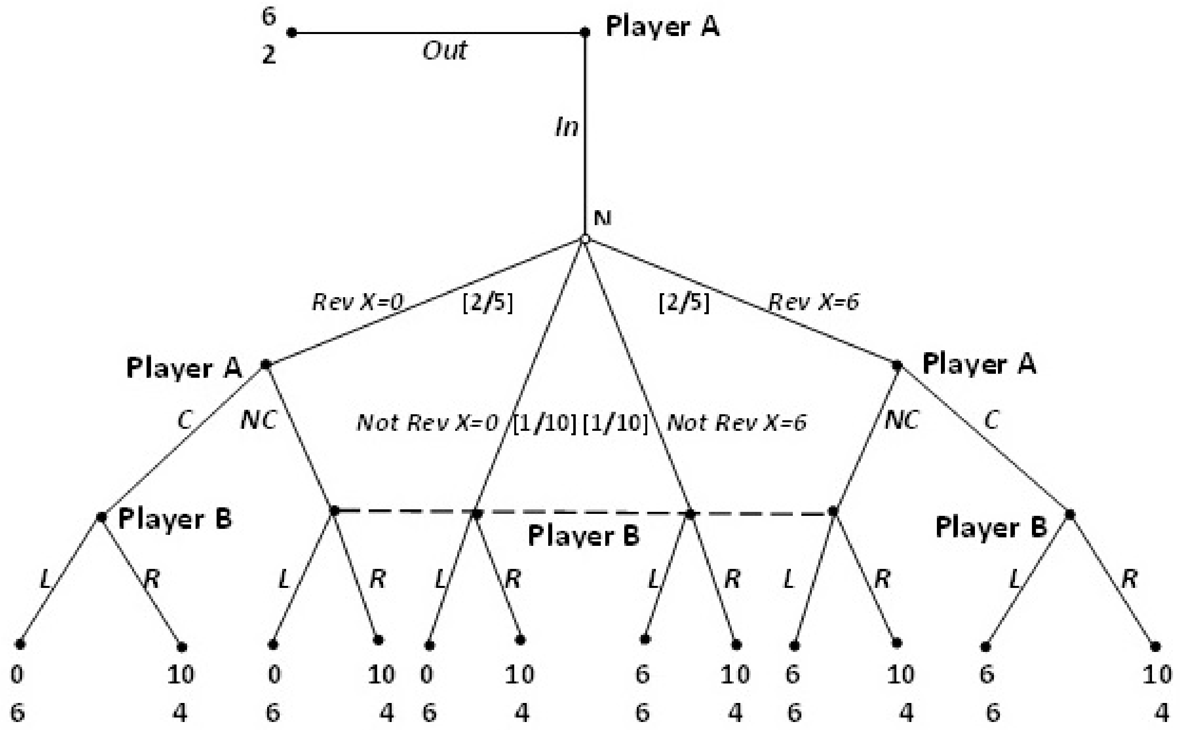
[10] Namely, a game where the first mover has an opportunity to choose an action that creates the possibility of mutual benefit if the other person cooperates, but a risk of lower payoffs to oneself if the other person defects. Such an action taken by the first mover is consistent with the behavioral definitions of trust presented in Cox (2004) and Fehr (2009).

Private Payoff Trust Game – Γ_{PPT}

Γ_{PPT} features a similar strategic structure and payoff structure to those of Γ_{UPT}, with two important differences. First, Γ_{PPT} features an opportunity for Player A to become privately informed about the value of X. Second, Γ_{PPT} features an additional stage where Player A has the opportunity to credibly convey his/her private information about the value of X to Player B. Γ_{PPT} begins analogously to Γ_{UPT} with Player A first choosing between *In* or *Out.* If Player A chooses *Out*, the game ends; Player A receives a payoff of 6, and Player B receives a payoff of 2. If Player A chooses *In,* Nature then decides whether Player A becomes privately informed about the value of X. With $prob = 4/5$, Nature *Reveals (Rev)* the value of X to Player A, and with $prob = 1/5$, Nature does *Not Reveal (Not Rev)* the value of X to Player A.

If the value of X is revealed to Player A, then an additional stage arises where Player A must decide whether to credibly *Convey (C)* on *Not Convey (NC)* the value of X to Player B *before* Player B gets the move. Upon getting the move, Player B must then decide between *Left* or *Right.* Analogous to Γ_{UPT}, if Player B chooses *Right,* then the game ends; Player A receives a payoff of 10, and Player B receives a payoff of 4. If Player B chooses *Left,* the game ends; Player A receives a payoff of X, and Player B receives a payoff of 6, where, again, $prob(X = 0) = 1/2$ and the $prob(X = 6) = 1/2$. The extensive form of Γ_{PPT} is depicted in Figure 2. To simplify the extensive form, the two moves by Nature — determining the value of X and determining whether the value of X is revealed to Player A — have been combined into one move.

Figure 2: Extensive Form of Γ_{PPT}



If players are "selfish", i.e., act to maximize their own material payoff, then the unique equilibrium outcome in both Γ_{UPT} and Γ_{PPT} is Player A chooses *Out.*[11] The inclusion of private information and the additional conveyance stage for Player A in Γ_{PPT} has no impact on the equilibrium outcome assuming selfish players. However, it is exactly these two features of Γ_{PPT} that will allow me to derive testable hypotheses regarding whether agents attempt to induce guilt and its subsequent effectiveness.

Before I proceed, I first highlight two important features of Γ_{PPT}, and discuss the motivation for including these features, which will be relevant for the upcoming derivation of the research hypotheses and the application of the B&D model of guilt. First, Player A must choose between *In* or *Out* before possibly becoming informed about the value of X (and Player B is informed of this timing). Not informing Player A of the value of X prior to the *In/Out* decision eliminates any possible signaling value, from the perspective of Player B, regarding the value of X inferred from Player A's *In/Out* decision. The second feature, is that the value of X is only revealed to a Player A who chooses *In* with $prob = 4/5$. The motivation for including this uncertainty regarding the revelation of X to Player A is the following: If the value of X is not conveyed to Player B, then Player B is unable to *perfectly* distinguish between whether: (i) the value of X was not revealed to Player A, or (ii) the value of X was revealed to Player A, and Player A chose to *Not Convey.* The resulting implication is that if the value of X is not conveyed to Player B, then Player B's expectation of the value of X will be *strictly* greater than zero and *strictly* less than six; the strict inequalities play an important role in the hypothesis development, discussed in Section 2.2 below.

More precisely, $\widehat{m}_A \in [1, 5]$ where $\widehat{m}_A$ denotes Player B's expectation of X, conditional on the value of X not being conveyed.[12] In calculating $\widehat{m}_A$, Player B must think about the relative probabilities that: (i) Player A learned $X = 0$ and chose *Not Convey,* (ii) Player A learned $X = 6$ and chose *Not Convey*, and (iii) Player A did not learn the value of X. Although these probabilities are unobservable to the researcher and, thus, the exact value of $\widehat{m}_A$ is unobservable, it is possible to derive bounds on $\widehat{m}_A$. Specifically, the largest expectation that Player B could hold regarding the value of X occurs when he/she thinks that only a Player A who learned that $X = 6$ would choose to *Not Convey.* In this case, $\widehat{m}_A$ is bounded above by $\frac{1}{3} \cdot E[X] + \frac{2}{3} \cdot 6 = 5$. Here, $\frac{1}{3}$ and $\frac{2}{3}$ represent the updated probabilities, via Bayes' rule, that Player A did not learn the value of X, and Player A learned that the value of $X = 6$, respectively. By a similar argument, the smallest expectation that Player B could hold regarding the value of X occurs when he/she thinks only a Player A who learned $X = 0$ would choose to *Not Convey.* In this case, $\widehat{m}_A$ is bounded below by $\frac{1}{3} \cdot E[X] + \frac{2}{3} \cdot 0 = 1$. Therefore, regardless of Player B's beliefs at his/her information set where the value of X is not conveyed, it must be that $\widehat{m}_A \in [1, 5]$.

2.2 Research Hypotheses

The first motivation of this study is to investigate whether agents attempt to exploit the guilt aversion of others by inducing guilt upon them. Recall that BSH (1994) posit that a person can induce guilt upon another by conveying to that person how one suffers over that person's failure to act in the desired fashion. Let us consider how this method applies to Γ_{PPT}. Conditional on

[11] In Γ_{PPT}, there is a multiplicity of Perfect Baysian Equilibria for selfish players that depend on the specification of Player B's beliefs at the information set where no information is conveyed regarding the value of X. However, regardless of Player B's beliefs, it is rational for him/her to choose *Left* and subsequently, it is sequentially rational for Player A to choose *Out.* Therefore, the unique equilibrium outcome of Γ_{PPT} is the game ending with Player A choosing *Out.*

[12] The notation of $\widehat{m}_A$ is consistent with the notation used in the B&D model of guilt aversion and anticipates the upcoming application of the model to Γ_{PPT}.

choosing *In*, Player A would "desire" Player B to choose *Right,* as it yields him/her a payoff of 10 compared to a payoff of $X < 10$ if Player B were to choose *Left.* Hence, X measures the extent to which Player A would "suffer" from Player B's failure to choose *Right.* Given the opportunity, Player A could attempt to induce guilt upon Player B by "conveying" to Player B a *low* value of X, i.e., by conveying to Player B how much he/she would suffer if Player B were to choose *Left.*

Conditional on having chosen *In* in Γ_{PPT} and having the value of X revealed, Player A must decide whether to credibly convey the value of X to Player B (i.e., Player A cannot lie about the value of X). In the case where $X = 0$ is revealed, if Player A chooses to *Convey* $X = 0$, then Player B will know that if he/she chooses *Left,* Player A will receive a payoff of $X = 0$. Whereas, if Player A chooses to *Not Convey* $X = 0$, then Player B will think that if he/she chooses *Left*, Player A will receive a payoff of $\widehat{m}_A \in [1, 5]$. Analogously, in the case where the value of $X = 6$ is revealed, if Player A chooses to *Convey* $X = 6$, then Player B will know that if he/she choose *Left,* Player A will receive a payoff of $X = 6$. Whereas, if Player A chooses to *Not Convey* $X = 6$, Player B will think that if he/she chooses *Left,* Player A will receive a payoff of $\widehat{m}_A \in [1, 5]$. Hence, from the perspective of Player B, Player A suffers *strictly* more from Player B's choice of *Left* when $X = 0$ is conveyed, compared to when the value of X is not conveyed. Similarly, from the perspective of Player B, Player A suffers *strictly* more from Player B's choice of *Left* if the value of X is not conveyed, compared to if $X = 6$ is conveyed.[13] Therefore, a Player A who is attempting to induce guilt upon Player B would *Convey* $X = 0$, and *Not Convey* $X = 6$. This leads to the first testable hypothesis:

H1: The proportion of Player As who *Convey* $X = 0$ in Γ_{PPT} is larger than the proportion of Player As who *Convey* $X = 6$.

The second motivation of this study is to investigate whether agents are susceptible to guilt induction. That is, are agents more motivated to respond kindly after guilt has been induced upon them? BSH (1994) posit that after Person A has induced guilt upon Person B, "Person B finds the guilt aversive and, to escape from guilt, complies with A's wishes" (p. 247). It is also possible, however, that a guilt averse Player B will recognize that Player A is trying to manipulate his/her behavior by "guilting" him/her, which can result in Player B being more motivated to choose the unkind action of *Left* in response to Player A's attempted guilt induction. BSH (1994, 1995) document this potential "cost" of guilt induction by arguing that the target of guilt induction (Player B) might feel resentment and be motivated to respond negatively toward the guilt inducer (Player A). Attempted guilt induction by Player A may actually be counterproductive as it may foster more selfish behavior and motivate Player B to choose *Left*, contrary to Player A's intended motivation. Hence, the susceptibility of agents to the exploitation of their guilt aversion, through guilt induction by others, is an open empirical question. If guilt induction by Player A is an effective influence mechanism, then Player B would be more motivated to choose *Right* after Player A induces guilt by choosing to *Convey* $X = 0$ and *Not Convey* $X = 6$. This leads to the following testable hypothesis:

[13] Note, if all Player As (who chose *In*) were revealed the value of X (as opposed to only 80% of them), then it would be possible, even probable, for Player B's belief about the value of X when it is not conveyed to be $X = 6$. Such a belief would be consistent with Player B thinking that *only* a Player A for whom $X = 6$ would choose *Not Convey.* Under this belief structure, the degree to which Player A suffers from Player B's choice of *Left*, from Player B's perspective, would be equal when $X = 6$ is conveyed and when X is not conveyed. Furthermore, if Player A anticipated beliefs of this sort from Player B, then Player A would be indifferent between conveying and not conveying $X = 6$; at least as far as attempted guilt induction is concerned. Hence, the inclusion of the move by Nature to reveal the value of X to Player A with $\textit{prob} = 4/5$ ensures that Player As who want to attempt to induce guilt upon Player B are motivated to *Not Convey* $X = 6$ (according to the BSH 1194 method).

H2: The proportion of Player Bs choosing *Right* in Γ_{PPT} after $X = 0$ is conveyed is larger than when the value of X is not conveyed, which is larger than when $X = 6$ is conveyed.[14]

The third motivation of this study is to investigate whether having the opportunity to induce guilt fosters more trusting behavior. Comparing Γ_{UPT} and Γ_{PPT}, we can see that the differences between Γ_{UPT} and Γ_{PPT} are the possibility for Player A to become privately informed about the value of X, and the ability to convey the learned value of X to Player B. As I have shown, it is these features of Γ_{PPT} that provide an opportunity for Player A to induce guilt upon Player B. Therefore, if having an opportunity to induce guilt fosters more trusting behavior, then Player As would be more motivated to choose *In* in Γ_{PPT}, compared to Γ_{UPT}. This leads to the following testable hypothesis:

H3: The proportion of Player As who choose *In* when playing Γ_{PPT} is larger than the proportion of Player As who choose *In* when playing Γ_{UPT}.

2.3 Consistency with the B&D Model of Guilt

Derived from how BSH (1994) posit that people induce guilt upon others, the way in which Player A would attempt to induce guilt upon Player B in Γ_{PPT} is by choosing to *Convey* $X = 0$ and *Not Convey* $X = 6$. I take this time to discuss how this manner of inducing guilt upon others is consistent with the B&D model of simple guilt. Namely, the B&D model would predict that Player B would suffer more guilt from choosing *Left* when $X = 0$ was conveyed, compared to when the value of X was not conveyed, compared to when $X = 6$ was conveyed. In what follows, I provide a brief explanation of why this is the case, and I refer readers to Appendix A, where a more formal derivation is provided including the application of the B&D model to Γ_{PPT}.

Recall that the B&D model posits that agents suffer disutility from guilt when they let down another relative to that other agent's expectations. Hence, in Γ_{PPT}, the amount of guilt that Player B will suffer from choosing *Left* will be proportional to the difference between the payoff that Player A was expecting to receive and the payoff that Player A *actually* receives as a result of Player B choosing *Left.* Conditional on Player A's payoff expectation when choosing *In*, the lower the actual payoff Player A will receive from Player B choosing *Left,* the more guilt Player B will suffer from choosing *Left.* By the structure of Γ_{PPT}, the lower the value of X the lower is Player A's actual payoff when Player B chooses *Left.* It follows that the B&D model would predict that Player B would feel more guilt from choosing *Left* when $X = 0$ was conveyed, compared to when the value of X was not conveyed, compared to when $X = 6$ was conveyed. Therefore, Player A can attempt to induce guilt upon Player B in Γ_{PPT} by choosing to *Convey* $X = 0$ and *Not Convey* $X = 6$. The idea is that because of the structure of Γ_{PPT} Player A can manipulate the guilt that Player B would feel via strategic revelation of private information to Player B about the *actual payoff* that Player A would receive if Player B chose *Left.*

The primary motivation of this study is to experimentally investigate whether agents attempt to exploit the guilt aversion of others, and whether agents are susceptible to such exploitation. These are questions related to behavioral motivations in games that do not depend upon any equilibrium supposition. Nevertheless, it is important to think about whether such behavior *can be* supported

[14] Note, if all Player As attempt to induce guilt, then $X = 6$ would never actually be conveyed to Player B; thus, no data would be generated on the proportion on Player Bs choosing *Right* after $X = 6$ was conveyed. In this case, H2 would just reduce down to the binary comparison of the proportion of Player Bs choosing *Right* after $X = 0$ was conveyed to when the value of X was not conveyed. However, as we will see in the Results section, some Player As do convey $X = 6$, so the necessary data is generated to test H2 as it is stated.

in equilibrium, as this is informative for determining whether such behavior is sustainable. It is the case that the two possible strategy profiles corresponding to guilt induction by Player A and a kind response to guilt induction by Player B can be supported as sequential equilibria of Γ_{PPT} under the guilt framework of B&D and the assumption of complete information about guilt sensitivities (refer to Appendix A for a more formal derivation of this). I acknowledge that assuming equilibrium play, especially when a game features multiple equilibria, is a rather strong notion; however, an equilibrium supposition is sufficient, and not necessary, for H1 and H2 to be consistent with predictions of the B&D model of guilt.

2.4 Possible Alternative Motivations

As shown above, Γ_{PPT} provides Player A an opportunity to induce guilt upon Player B and, thus, a way of exploring whether Player As attempt to do so and whether Player Bs are susceptible to such an attempt by Player As. In general, however, the behavior of Player Bs may be influenced by other factors besides guilt aversion. I conclude this section by briefly discussing some alternative motivations that could possibly be impacting the behavior of Player Bs in Γ_{PPT} and how these relate to the research hypotheses.

One possible motivation is inequality aversion — the idea that agents are averse to unequal outcomes, both advantageous and disadvantageous (e.g., Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000 for seminal models). Applied to trust games, inequality aversion is often cited as a possible explanation for why second movers may by motivated to choose the more kind action (*Right*), as it generally results in a more equal division of payoffs. However, the payoffs in Γ_{PPT} were designed to attempt to control for possible inequality aversion of Player Bs. Specifically the payoffs are structured such that, from the perspective of Player B, a choice of *Left* in Γ_{PPT}, compared to a choice of *Right,* yields: (i) a higher material payoff, (ii) weakly less inequality in the payoff distribution between Player A and B (in terms of the absolute difference in payoffs), and (iii) favorable rather than unfavorable payoff inequality for all values of $X \in [0, 6]$. Given these three properties, the Fehr and Schmidt model, and the nonlinear extension proposed by Bellemare et al. (2008), would predict that inequality averse Player Bs would always be motivated to choose *Left,* regardless of the value of X.[15] Thus, as specified by the linear and nonlinear versions of the Fehr and Schmidt model, behavior by Player B in the direction of H2 cannot be explained by possible inequality aversion of Player B; hence, variation in Player A's conveyance decision regarding the value of X in Γ_{PPT} cannot be explained by any strategic considerations regarding inequality aversion of Player B as characterized by the Fehr and Schmidt model.

I acknowledge, however, that the parameterization of Γ_{PPT} does not fully control for inequality averse preferences as specified by the Bolton and Ockenfels (2000) model. In their model, inequality is measured based on a player's *relative* share of his/her payoff (i.e., a player's payoff divided by the total payoff across all players); subsequently, players *dislike* when their payoff deviates away from the equal, average relative share of the payoffs, which in a 2-player game is simply when a player's relative payoff share is 1/2. In the context of Γ_{PPT}, this implies that it is possible for a Player

[15] To illustrate, consider the application of the Fehr and Schmidt (1999) model to Player B in Γ_{PPT}. An inequality averse Player B playing Γ_{PPT} always prefers the payoff vector $(X, 6)$ to the payoff vector $(10, 4)$ $\forall X \in [0, 6]$. To see this, note that for the extreme case where $X = 0$, we have that $6 - \beta(6 - 0) > 4 - \alpha(10 - 4)$ $\forall \beta \in [0, 1)$ and $\alpha \geq \beta$, where the LHS represents an inequality averse Player B's utility from choosing *Left* and the RHS represents the utility from choosing *Right.* For the other extreme case where $X = 6$, we have that $6 > 4 - \alpha(10 - 4)$ $\forall \alpha \geq 0$, where the LHS represents an inequality averse Player B's utility from choosing *Left* and the RHS represnts the utiility from choosing *Right.* The constraints that $\beta \in [0, 1)$ and $\alpha \geq \beta$ in the above inequalities are assumed a priori in the FS model.

B, with Bolton and Ockenfels type inequality averse preferences, to prefer *Right* when $X = 0$. Specifically, it is possible (under the very general assumptions of the model) that Player B would prefer choosing *Right* and sacrificing \$2 in own payoff (moving from \$6 to \$4), in order to bring his/her relative payoff share closer to 1/2 (moving from $6/6 = 1$ to $4/14 = 2/7$). That said, such a choice would require Player B sacrificing 33% of his/her own material payoff, as well as moving to a position of unfavorable/disadvantageous inequality, to move 28% closer to the relative payoff share of 1/2. Moreover, Englemann and Strobel (2004) provide experimental evidence that the Fehr and Schmidt model better predicts behavior than does the Bolton and Ockenfels model. In summary, the payoffs in Γ_{PPT} were structured to control for possible inequality aversion of Player Bs as posited by the Fehr and Schmidt model. However, I cannot go as far as saying the design fully controls for all possible characterizations of preferences for inequality aversion. It is possible, although unlikely in my view, that a Player B may be more motivated to choose *Right* when $X = 0$ (the direction of H2) because it moves Player B's relative payoff share closer to 1/2.[16]

Another possible motivation for Player Bs may be positive reciprocity — the idea that agents may be motivated to respond kindly to agents who are kind to them (e.g., Dufwenberg and Kirchsteiger, 2004). In the context of Γ_{PPT}, Player B may be motivated to positively reciprocate the kind action of Player A choosing *In* by choosing *Right,* and this may be more prevalent if Player B knows that $X = 0$ (see Cox, 2004 for a general discussion of reciprocity motivations in trust games). To control for this possible reciprocity confound, I consider a third game, which is a "dictator" version of Γ_{PPT}. In this modified dictator version of Γ_{PPT}, denoted as Γ_{PPD}, the initial *In/Out* decision of Player A is eliminated. Γ_{PPD} begins with Player A's decision to convey X to Player B, conditional on X being revealed, and proceeds with Player B's decision of *Right* or *Left.* Hence, the extensive form of Γ_{PPD} is simply the subgame of Γ_{PPT} that begins with Nature's move. The removal of the initial *In/Out* decision eliminates possible motivations to positively reciprocate Player A's *In* decision from Player B's *Right* decision. If there are no significant differences between Player B behavior in Γ_{PPD} and Γ_{PPT}, then reciprocity motivations are not a salient concern, and H2 can be tested using pooled data from both Γ_{PPT} and Γ_{PPD}. However, if Player B behavior differs across the two games, then reciprocity could be salient, and H2 will be tested using data from Γ_{PPD} only. By comparing the decision making of Player B in Γ_{PPT} with decision making in the dictator version of Γ_{PPT}, I am able to test for and, if necessary, control for possible reciprocity motivations of Player B.[17]

However, one limitation of the experimental design is its inability to directly control for efficiency concerns or maximin preferences (Charness and Rabin, 2002), both of which may be salient motivations (see Engelmann and Strobel, 2004). In regards to efficiency, Player B may be motivated to make choices that maximize the sum of total payoffs. Note that in Γ_{PPT}, *Right* is always more efficient, so in the extreme case where Player B was solely motivated, or very strongly motivated by efficiency concerns, then Player B would always be motivated to choose *Right.* However, under the more reasonable assumption that efficiency concerns are being balanced against one's own material payoffs, then at the margin Player B may be more motivated to choose *Right* after $X = 0$ had been

[16] In my view it is even much less plausible that this could confound the inference regarding Player A's attempted guilt induction of Player B (H1). Specifically, this argument would require Player A to somehow anticipate that Player B had inequality averse preferences based on relative payoffs shares (e.g., the Bolton and Ockenfels model), as well as then reason through that conveying $X = 0$ induces a payoff vector where such an inequality averse Player B would then be more motivated to choose *Right.* That being said, I do acknowledge that this is a possible confound and shortcoming of the experimental design.

[17] This approach of controlling for the possible reciprocity concerns of Player B in the trust game (Γ_{PPT}) by using the corresponding dictator game (Γ_{PPD}) was inspired by the triadic design approach developed and implemented in Cox (2004).

conveyed. Similarly, if Player B has maximin preferences — a desire to maximize the lowest payoff of a player in the game — then Player B may be more motivated to choose *Right* after $X = 0$ had been conveyed.

It is important to note that from a conceptual standpoint, it may be difficult to eliminate concerns for efficiency and maximin preferences when testing if agents are susceptible to guilt induction. The reason rests in the manner by which agents induce guilt upon others, namely, by conveying that they will receive a low payoff given an undesirable action of the other agent. Therefore, relative to not conveying such information, the undesirable choice will necessarily lead to a lower payoff for the other agent and a lower level of efficiency, as is the case in Γ_{PPT}. As a result, I am unable to separate out efficiency concerns and maximin preferences from guilt aversion when testing whether Player Bs are susceptible to guilt induction by Player As (H2).

That said, Γ_{PPT} allows Player A to induce guilt upon Player B by conveying $X = 0$ and not conveying $X = 6$, which is consistent with the method posited by BSH (1994). Furthermore, the B&D model predicts that a Player B would suffer more guilt from choosing *Left* after $X = 0$ is conveyed, compared to when the value of X is not conveyed, compared to when $X = 6$ is conveyed. Hence, if the experimental data supports H1, then I will interpret this as evidence consistent with Player As attempting to exploit the guilt aversion of Player Bs. Subsequently, if the experimental data supports H2, then I will interpret this as evidence consistent with Player Bs being susceptible to the exploitation of their guilt aversion. Readers should certainly be mindful of the fact that other possible motivations, e.g., efficiency concerns and/or maximin preferences, could be influencing the behavior of Player B in addition to guilt aversion. However, in the results section, I provide some results from a post-decision questionnaire that are consistent with Player Bs being motivated, at least in part, by guilt aversion. This is consistent with recent findings by Attanasi et al. (2014b) who document that "guilt aversion is a prevelent psychological motivation" (p. 3) in a 2-player trust game.

3 Experimental Procedure

All experimental sessions were conducted in the Economic Science Laboratory (ESL) at the University of Arizona in April 2011 and October 2011. The sessions were computerized, and the software was programmed using Z-tree (Fischbacher, 2007). The subject pool consisted of undergraduates who were recruited via an online database. In total, 22 sessions were conducted using 444 subjects comprising 222 two-player groups.

To test the three main research hypotheses of this paper (H1-H3), I use a between-groups design where all participants are randomly assigned to one of the following three experimental treatments:

UPT Treatment Subjects played Γ_{UPT}, where the payoffs from Γ_{UPT} corresponded 1:1 with the monetary payoffs in the experiment.

PPT Treatment Subjects played Γ_{PPT}, where the payoffs from Γ_{PPT} corresponded 1:1 with the monetary payoffs in the experiment.

PPD Treatment Subjects played Γ_{PPD}, where the payoffs from Γ_{PPD} corresponded 1:1 with the monetary payoffs in the experiment

Conditional on their random assignment to treatment, all participants were randomly assigned to either the role of Player A or Player B and then randomly and anonymously matched with a participant of the opposite player role. All participants then proceeded to play their designated game *one time* in their assigned player role. Of the 222 groups, 111 were assigned to the PPT Treatment, 45 were assigned to the UPT Treatment, and 66 were assigned the PPD Treatment. Each session lasted approximately 25 minutes and subjects earned an average of $9.68 USD (including a $5 show-up payment).

Before starting the game, the experimental instructions were carefully read aloud to all participants in the session to enhance clarity and general understanding of the task among the participants. A copy of the experimental instructions can be found in Appendix B.1, as well as sample screen shots of the computer interface in Appendix B.2. Upon the completion of the game, the decisions of each player, the corresponding outcome, the profit to each player, and the value of X were displayed to both players. All subjects were informed in the instructions that the value of X would be revealed to both players upon completion of the task, irrespective of the decisions made in the task. Revealing the value of X to all players ensures that Player As were not motivated to choose *In* (or *Conveying* X) just so Player A (Player B) could learn the value of X. This design feature eliminates any curiosity biases that may arise from the uncertainty regarding the value of X, and the consequent payoffs to the other player.

Upon completion of the game, subjects were asked to fill out a short questionnaire. In all treatments, the questionnaire contained 8 general demographic questions. In the PPT and PPD treatments, two additional questions were asked that related to guilt feelings and perceptions of guilt feelings in the game. These specific guilt related questions, the corresponding responses, and discussion of the possible gleaned insights from these questions are presented in Section 4.2.

4 Results

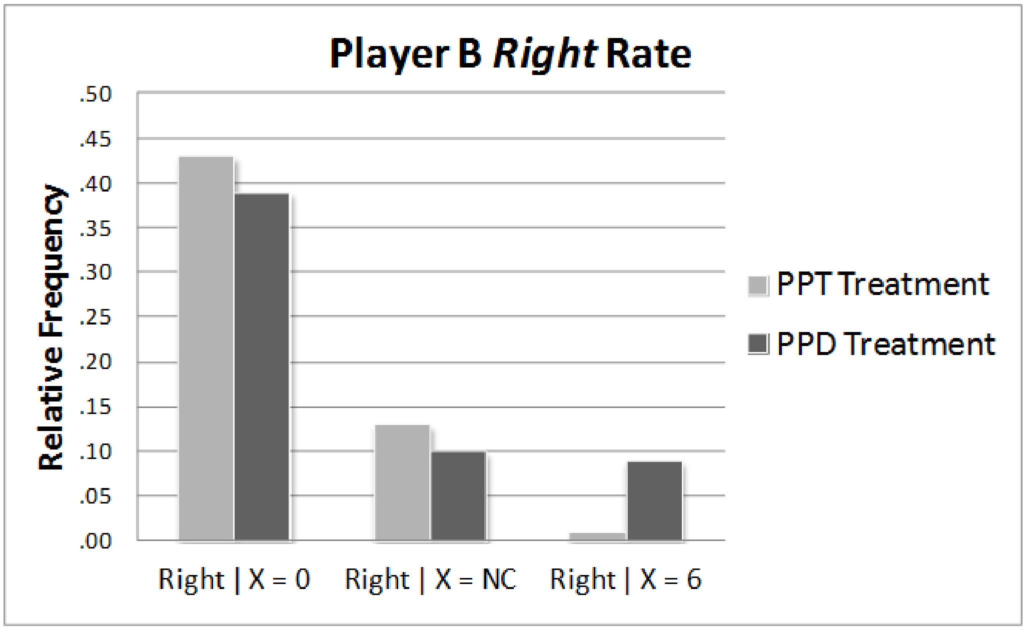
I first present the aggregate decision data from each of the three treatments and the corresponding tests of the three research hypotheses. I then present some results from the post-decision questionnaire. I conclude with discussion and speculative remarks about some of the observed patterns in the data.

4.1 Aggregate Data and Hypothesis Testing

I begin by comparing the aggregate Player B data from the PPT and PPD treatments in order to test for any possible reciprocity motivations (see Section 2.4). Specifically, I compare the frequency of Player Bs choosing *Right* at each of the three possible conveyance states: (i) $X = 0$ was conveyed (Right$|X = 0$), (ii) the value of X was not conveyed (Right$|X = NC$), and (iii) $X = 6$ was conveyed (Right$|X = 6$). Figure 3 displays the histogram of the relevant Player B data.

From Figure 3, we can see that the relative frequencies of *Right* choices at each of the three conveyance states are similar across the two treatments. In fact, a 2-sided Fisher's Exact test does not yield a significant difference between the proportion of Player Bs who choose *Right*$|X = 0$, *Right*$|X = NC$, and *Right*$|X = 6$ between the two treatments ($p = 1.000$, $p = 1.000$, and $p = 0.542$, respectively). Thus, we can rule out reciprocity as a salient confounding motivation for Player B's *Right/Left* decision in Γ_{PPT}, conditional on the conveyance state. As a result, I am able to proceed in testing H2 using pooled data from the PPT and PPD treatments, as discussed in Section 2.4, which provides the added benefit of a larger sample size and more statistical power for the hypothesis testing. That said, because of differences in the strategic structure of the two games (dictator vs trust), I will also present the results and hypothesis testing separately for the PPT and PPD treatments.

Figure 3: Comparison of *Right* Rates – PPD and PPT Treatments



I test the hypotheses in the order that corresponds to working backwards through the game. Namely, I first test H2, whether Player Bs are susceptible to guilt induction by Player As. In terms of testing H2, I compare the proportion of Player Bs who choose $Right|X = 0$, $Right|X = NC$, and $Right|X = 6$. Table 1 presents the relevant data from the Player Bs who were called upon to make a decision (i.e., all 66 Player Bs from the PPD treatment and the 58 Player Bs in the PPT treatment whose corresponding Player A chose *In*). From Table 1, we can see that in the PPT treatment 3/7 (43%) Player Bs chose $Right|X = 0$, 5/40 (13%) chose $Right|X = NC$, and 0/11 (0%) chose $Right|X = 6$. The corresponding frequencies in the PPD treatment are 9/23 (39%), 2/21 (11%), and 2/22 (10%). A Jonckheere-Terpstra non-parametric test for ordered alternatives rejects the null of equality of these *Right* rates in favor of the ordered alternative for each treatment ($p = 0.008$ and $p = 0.006$, respectively), as well as pooled over both treatments ($p < 0.001$).[18] Thus, the data suggests that Player Bs are susceptible to guilt induction by Player As, in the sense that Player Bs are more likely to choose $Right|X = 0$, compared to $Right|X = NC$, compared to $Right|X = 6$, which supports H2. An alternative interpretation is that guilt induction by Player A *would be* effective at increasing the likelihood that Player B chooses *Right*; this is especially true when explicitly choosing to convey $X = 0$, rather than not conveying $X = 0$. The non-linear pattern in *Right* rates across these three conveyance states likely results from Player B's belief about the value of X in the NC state, $\widehat{m}_A$, not being equal to 3. Specifically, the data suggests that $\widehat{m}_A$ is close to its upper bound of 5, which corresponds to the case where Player B believes that only those Player As for whom $X = 6$ would not convey.

[18] Alternatively, a simple probit regression, with *Right* as the dependent variable and the conveyance state as the independent variable (where $X = 6$ is coded as a zero, $X = NC$ is coded as a one, and $X = 0$ is coded as a two) also yields a significantly positive coefficient ($p = 0.001$) on the conveyance state variable.

Table 1: Comparison of Player B *Right* Rates by Conveyance State

	Aggregate Player B *Right* Rate			
Treatment	$Right \mid X = 0$	$Right \mid X = NC$	$Right \mid X = 6$	(p-value)
PPT Treatment	3/7 (43%)	5/40 (13%)	0/11 (0%)	(.008)
PPD Treatment	9/23 (39%)	2/21 (11%)	2/22 (10%)	(.006)
Pooled over PPT and PPD	12/30 (40%)	7/61 (11%)	2/33 (6%)	(< 0.001)

Notes: reported p-values are from a Jonckheere-Terpstra non-parametric test for ordered alternatives.

Next, I turn to testing H1. Namely, are Player As attempting to exploit the guilt aversion of Player Bs by choosing to *Convey* $X = 0$ more frequently than they choose to *Convey* $X = 6$? Table 2 presents the relevant conveyance data from the Player As who were called upon to make a conveyance decision (i.e., 56 Player As from the PPD treatment to whom X was revealed, and 42 Player As from the PPT treatment who chose *In* and to whom X was revealed). From Table 2 we can see that in the PPT treatment 7/16 (44%) Player As chose to *Convey* $X = 0$ and 11/26 (42%) Player As chose to *Convey* $X = 6$. The corresponding frequencies in the PPD treatment are 23/27 (85%) and 22/29 (76%), respectively. Clearly, the data reveals that in neither treatment did *all* Player As choose to *Convey* $X = 0$ and *Not Convey* $X = 6$. Furthermore, while the proportion of Player As who *Convey* $X = 0$ is larger than the proportion who *Convey* $X = 6$ in both treatments, the difference is not significant, using a 1-sided Fisher's exact test, in either treatment ($p = 0.589$ and $p = 0.296$ respectively); likewise, the difference in conveyance rates is not significant for the pooled sample using a 1-sided Fisher's exact test ($p = 0.215$) or a 1-sided t-test ($p = 0.161$). Looking separately at the PPT and PPD treatments, or pooled over both treatments, the Player A conveyance data fails to support H1, which suggests that Player As are not attempting to fully exploit the guilt aversion of Player Bs by inducing guilt upon them.

Before presenting the aggregate trust rates in the PPT and UPT treatments and the corresponding test of H3, I pause to briefly discuss the differences in Player A conveyance rates across the PPT and PPD treatments. Specifically, the overall rate of conveyance is significantly higher in the PPD treatment compared to the PPT treatment, both for $X = 0$ and $X = 6$. Given the differences in the strategic structure across the games (i.e., the presence of the outside option in PPT along with the initial *In/Out* decision for Player A), there is no reason to assume, ex-ante, that the conveyance behavior be equal across the two games.[19] That said, the identification of

[19] Some factors that may have contributed to the higher levels of conveyance in the PPD treatment include, but are not limited to: (i) lack of the selection effect that could have been present in the PPT treatment, since all Player As were forced to make a conveyance decision in PPD, (ii) that inability to signal trust in PPD, which may have prompted Player As to be more likely to attempt to signal honesty by conveying X in hopes of increasing the likelihood of Player B choosing the kind allocation, and/or (iii) activity bias/experimenter demand effects for Player As in the PPD treatment.

attempted guilt induction by Player A (H1) is based on the relative comparison of the ratio of Player As who convey $X = 0$ with the ratio of those who convey $X = 6$. Thus, the level differences in the conveyance rates across the two games do not necessarily invalidate the test of H1 within each treatment, or the overall implication drawn from the data in the previous paragraph that Player As do not appear to be attempting to induce guilt upon Player Bs.

Table 2: Comparison of Player A Conveyance Rates by Value of X

	Aggregate Player A Data		
Treatment	*Convey* $X = 0$	*Convey* $X = 6$	(p-value)
PPT Treatment	7/16 (44%)	11/26 (42%)	(0.589)
PPD Treatment	23/27 (85%)	22/29 (76%)	(0.296)
Pooled over PPT and PPD	30/43 (70%)	33/55 (60%)	(0.215)

Notes: p-values are reported for a 1-sided Fisher's Exact test of conveyance rates

Lastly, I turn to testing H3; namely, does having an opportunity to induce guilt foster more trusting behavior? To test this hypothesis, I compare the aggregate *In* rate from the 111 Player As in the PPT Treatment with the *In* rate from the 45 Player As in the UPT Treatment. From Table 3, we can see that 58/111 (52%) of Player As chose *In* in the PPT Treatment and 17/45 (38%) of Player As chose *In* in the UPT treatment, which is significant using a 1-sided Fisher's Exact test ($p = 0.071$) and a 1-sided t-test ($p = 0.051$).[20] Thus, the aggregate data on Player A *In* rates across the PPT and UPT treatments marginally supports H3, which suggests that strategic settings that provide an opportunity for agents to induce guilt may foster more trusting behavior by those agents.

Before proceeding, I pause to acknowledge that the observed difference in *In* (trust) rates between the PPT and UPT games could result from alternative motivations besides Player As foreseeing an opportunity to induce guilt in PPT; this is especially so in light of the fact that the data reveals little evidence that Player As attempted to induce guilt. For example, some alternative explanations for why Player As in PPT may be more motivated to choose *In* could be: (i) they have an opportunity to communicate information, (ii) they can learn the value of X *before* Player B acts (as opposed to the end of the game in UPT), or (iii) Player A feels they have some increased ability to control the outcome in PPT.[21] Thus, a more conservative interpretation would be that the marginal support of H3 provides evidence that the inclusion of private payoff information for

[20] The 38% and 52% observed rates of trust are largely consistent with findings in prior studies that have implemented variations of binary-choice trust games. For example, Bohnet and Huck (2004) document a range of trust rates from 19% - 59%, Charness and Dufwenberg (2006) from 23% - 74%, Charness and Dufwenberg (2011) from 44% - 80%, and Charness et al. (2011) from 25% - 54%.

[21] I thank two anonymous reviewers for aptly calling attention to these possible alternative explanations.

Player A about their payoff if Player B is not trustworthy and the ability to convey this information to Player B can increase trust by Player A. That said, as I noted in the Introduction, these two features are necessary for providing Player A with the opportunity to induce guilt upon Player B.

Table 3: Comparison of Player A In Rates

	UPT Treatment	PPT Treatment	(p-value)
In Rate	17/45 (38%)	58/111 (52%)	(0.071)

Notes: p-value is reported from a 1-sided Fisher's Exact test of conveyance rates

4.2 Questionnaire Results

Next, I present results from two post-decision questionnaire questions asked to both Player As and Player Bs in the PPT and PPD treatments. The motivation of these two questions was to gain additional insights regarding Player B's feelings of guilt, and Player A's perceptions of Player B's feelings of guilt. This questionnaire was not incentivized and did not impact monetary earnings. As a result, the natural amount of discretion must be used in evaluating the resulting data. At the same time, there is really no scope for any type of monetary gains from strategic false reporting and, thus, no obvious material incentive to not report truthfully.[22] In addition, all the analysis of the response data is done using matched samples, which controls for scaling differences and possible anchoring effects that could exist across unmatched samples.

For Player As, the first question asked how much guilt they thought Player B felt (would have felt) from choosing *Left* if Player B knew (would have known) the true value of X. The second question asked Player As how much guilt they thought Player B felt (would have felt) from choosing *Left* if Player B did not know (would not have known) the true value of X. Responses were ranked on a 5-point scale with 5 being a *Very High* amount of guilt and 1 being a *Very Low* amount of guilt. For the analysis, I consider only the Player As who had the value of X revealed to them and, thus, had an opportunity to convey the value of X to Player B. Table 4 presents the aggregate response data for these Player As.[23] Table 4 is divided into two panels that correspond to whether $X = 0$ (Panel 1) or $X = 6$ (Panel 2) was revealed to Player A. Within each panel, the average reported perceptions of Player B's guilt feelings from choosing *Left* are separately presented for those Player As who conveyed the corresponding value of X and those who did not.

[22] I note, however, that agents may be motivated by non-material incentives to align their ex-post beliefs about guilt perceptions with their behavior. An example of such a motivation would be to avoid the disutility or discomfort associated with the inconsistency between beliefs and actions, a phenomenon that psychologists generally refer to as cognitive dissonance. Hence, the maintained disclaimer about interpreting these questionnaire results with caution.

[23] Because of the way the questionnaire was programmed in z-tree and administered, it was possible for subjects to fail to submit an answer for each question. A total of 5 of 98 Player As who had the value of X revealed to them failed to answer at least one of the questions related to their beliefs of Player B's guilt feelings. Therefore, the aggregate data in Table 4 reflects the responses of 93 Player As who did answer both questions.

Table 4: Player A's Perceptions of Player B's Guilt Feelings from Choosing Left

Panel 1 – Player As for whom $X = 0$ was Revealed

	Guilt if Player B knew $X = 0$	Guilt if Player B did not know X	(p-value)
Player As who conveyed $X = 0$	2.46	1.78	(0.008)
Player As who did not convey $X = 0$	2.36	1.91	(0.524)

Panel 2 –Player As for whom $X = 6$ was Revealed

	Guilt if Player B knew $X = 6$	Guilt if Player B did not know X	(p-value)
Player As who conveyed $X = 6$	1.72	2.21	(0.061)
Player As who did Not Convey $X = 6$	1.48	2.49	(0.002)

Notes: Reported p-values are from a Wilcoxon Signed-Rank test

From Panel 1 of Table 4, we can see that the Player As who chose to *Convey* $X = 0$ perceived that Player B would have felt significantly more guilt from choosing *Left*$|X = 0$ compared to *Left*$|X = NC$ ($p = 0.008$). However, the Player As who chose to *Not Convey* $X = 0$ did not perceive that Player B would have felt significantly more guilt from choosing *Left*$|X = 0$ compared to *Left*$|X = 0$ ($p = 0.524$). Similarly, from Panel 2, we see that the Player As who chose to *Not Convey* $X = 6$ perceived that Player B would have felt significantly more guilt from choosing *Left*$|X = NC$ compared to *Left*$|X = 6$ ($p = 0.002$). Yet, Player As who chose to *Convey* $X = 6$ perceived that Player B would have felt only marginally more guilt from choosing *Left*$|X = NC$ compared to *Left*$|X = 6$ ($p = 0.061$).[24]

[24] One seemingly inconsistent pattern that emerges from Table 4 is the comparison of the average level of reported guilt that Player A thought Player B would feel from choosing *Left* when "Player B did not know X" (column 2), depending on whether Player As were revealed $X = 0$ (Panel 1) or $X = 6$ (Panel 2). In the former case, the average reported level of guilt is approximately 1.85 (aggregated over Player As who did and did not convey X), which is significantly lower than in the latter case, where the average reported level of guilt is approximately 2.35. I speculate that this difference is the result of an anchor/adjustment process. Specifically, Player As for whom $X = 0$ report a high level of guilt that Player B would feel from choosing *Left* if $X = 0$ was known (the anchor point), and then adjust this level *downward* if X was unknown. Those Player As for whom $X = 6$ report a low level of guilt that Player B would feel from choosing *Left* if $X = 6$ was known (the anchor point), and then adjust this level *upward* if X was unknown. Thus, the difference is post-adjusted levels of reported guilt is likely a result of the initial anchoring level and the corresponding magnitude of the adjustment. That said, the analysis of the questionnaire data was

One of the primary things revealed from the Player A response data is that the Player As whose conveyance decisions were consistent with attempted guilt induction (those Player As who conveyed $X = 0$ and did not convey $X = 6$) seemed to think that doing so would induce more guilt upon Player B if Player B were to choose *Left.* However, the Player As whose conveyance decisions where not consistent with attempted guilt induction (those Player As who did not convey $X = 0$ and did convey $X = 6$) seemed to think that doing so would not have induced as much guilt upon Player B if Player B were to choose *Left.*

I turn now to the Player B questionnaire data. For Player Bs, the first question asked how much guilt he/she felt (would have felt) from choosing *Left* (if he/she had chosen *Left*) if he/she did not know the value of X. The second question asked Player B how much guilt he/she felt (would have felt) from choosing *Left* (if he/she had chosen *Left*) if he/she knew the value of X. Again, responses were ranked on a 5-point scale with 5 being a *Very High* amount of guilt and 1 being a *Very Low* amount of guilt. For the analysis, I consider the Player Bs who actually made a *Left/Right* decision in either Γ_{PPT} or Γ_{PPD}. Table 5 presents the aggregate response data for these Player Bs.[25] Table 5 is divided into two panels that correspond to the actual value X, and each panel shows the average reported Player B guilt feelings from choosing *Left.*

Table 5: Player B's Reported Guilt Feelings from Choosing Left

Panel 1 – Player Bs for whom $X = 0$		
Guilt if $X = 0$ Known	Guilt if X Unknown	(p-value)
2.53	1.73	(< 0.001)
Panel 2 – Player Bs for whom $X = 6$		
Guilt if $X = 6$ Known	Guilt if X Unknown	(p-value)
1.42	2.15	(< 0.001)

Notes: Reported p-values are from a Wilcoxon Signed-Rank test

From Panel 1 of Table 5 we see that when $X = 0$, Player Bs reported that they would have felt significantly more guilt from choosing *Left*$|X = 0$ compared to *Left*$|X = NC$ ($p < 0.001$). Similarly, from Panel 2 we see that when $X = 6$, Player Bs reported that they would have felt significantly more guilt from choosing *Left*$|X = NC$ compared to *Left*$|X = 6$ ($p < 0.001$). These reported perceived guilt feelings by Player Bs are consistent with guilt aversion acting as a predominant

done using a signrank test of matched data for each participant, and the identified effects are based on the *relative* differences in the participants' reported level of guilt across the two questions. As a result, possible anchoring effects that may distort the absolute level of reported guilt, do not invalidate the matched pairs analysis based on the relative comparison of guilt across questions.

[25] A total of 7 of 124 of the Player Bs did not answer at least one of the questions related to their guilt feelings. Therefore, the aggregate data in Table 5 reflects the responses of 117 Player Bs who did answer both questions.

motivation in the observed aggregate behavior of Player Bs being more likely to choose $Right|X = 0$, compared to $Right|X = NC$, compared to $Right|X = 6$. This reinforces the idea that Player Bs are susceptible to the exploitation of their guilt aversion by Player A.

4.3 Discussion

Although the motivation of this study is to test H1-H3, I take this time to make a few speculative remarks regarding some of the observed patterns in the data. I begin by proposing some plausible explanations of why Player As seem to not be fully exploiting the guilt aversion of Player Bs (i.e., failure to support H1). Failure by Player A in attempting to induce guilt upon Player B in Γ_{PPT} (and Γ_{PPD}) corresponds to Player As either (i) not conveying $X = 0$, and/or (ii) conveying $X = 6$. The data revealed a non-trival percentage of Player As did not convey $X = 0$ (30%) *and* did convey $X = 6$ (60%), which suggests that failure to support H1 seems to be a result of a combination of both (i) and (ii).

One possibility is that some Player As just don't realize or think that a sufficient amount of guilt could have been induced upon Player B by conveying $X = 0$ and not conveying $X = 6$. Consequently, these *types* of Player As would not necessarily be motivated to convey $X = 0$ and/or not convey $X = 6$ as a means of inducing guilt. The questionnaire data presented in the previous section provided some evidence consistent with this possibility; namely, we saw that the Player As who made conveyance decisions that were inconsistent with attempted guilt induction did not seem to think *as much* guilt could have been induced by conveying $X = 0$ and not conveying $X = 6$, compared to those Player As whose conveyance decisions were consistent with attempted guilt induction. A second possibility is that some Player As feel guilt over inducing guilt, an idea that BSH (1994) refer to as *metaguilt.* Player As who are averse enough to these metaguilty feelings would then be motivated to not induce guilt by not conveying $X = 0$ and conveying $X = 6$. A third possibility is that some Player As were attempting to "signal" honesty to Player B by not attempting to manipulate Player B's behavior by letting them think that $X = 0$, when in fact $X = 6$; said differently, some Player As who learned $X = 6$ might not have wanted to deceive Player B into thinking $X = 0$. The motivation for doing this is that Player As may have strategically reasoned that this display of honesty (as opposed to deception) could, in turn, increase B's trustworthiness and increase the likelihood of Player B choosing Right.[26] In fact, Wang et al. (2009) and Wang and Leung (2010) document recent experimental evidence of this effect where honesty (in the form of information revelation) is rewarded, while deception is punished; relatedly, Brandts and Charness (2003) document experimental evidence that agents are punished less frequently when honest.

Regarding the Player A conveyance data, it is important to note that the observed failure to support H1 is not evidence that agents universally would not attempt to induce guilt upon others across all strategic environments, and I would caution readers from interpreting it as such. Rather, the data from this study provides evidence that, at least in the specific strategic settings considered (the PPT and PPD games), participants acting in the role of Player A did not fully capitalize on their opportunity to induce guilt upon Player Bs; this is especially true with regard to conveying $X = 0$, where they could have increased the likelihood of Player B choosing *Right* by about 30% (in expectation). Amongst the set of Player As who thought guilt could be induced, their behavior seems to be in line with attempted guilt induction (Table 4). However, there is also evidence of a non-trivial fraction of Player As exhibiting behavior inconsistent with inducing guilt, which could

[26] An alternative, non-strategic, explanation is that Player As who viewed not conveying $X = 6$ as deceiving Player B (into thinking $X = 0$) could have been averse to such a deceptive act. Gneezy (2005), and Erat and Gneezy (2012) provide experimental evidence of deception aversion.

be for a myriad of reasons; a non-exhaustive list of some plausible reasons is listed above. Future research is warranted that aims at exploring the possible mediating and moderating factors of an agent's decision to attempt to induce guilt upon others in strategic decision making environments.

With regard to Player B behavior, a pattern that emerges from the data is Player Bs' propensity to choose *Right* is not linearly increasing across the conveyance states. Player Bs are much more likely to choose $Right|X = 0$ (40%) than $Right|X = NC$ (11%), as compared to $Right|X = 6$ (6%). Recall, BSH (1994) posit that guilt is induced by conveying how one suffers if another fails to act in a desired fashion. You can think of Player A choosing to convey $X = 0$ as an explicit attempt to induce guilt because Player A is actually conveying how he/she suffers. On the other hand, you can think of Player A choosing to not convey $X = 6$ as an implicit attempt to induce guilt because Player A is conveying how he/she suffers by not conveying how he/she doesn't suffer. This distinction might be important in terms of the effectiveness of guilt induction, in light of the fact that Player Bs were much more likely to choose *Right* after Player A had explicitly induced guilt by conveying $X = 0$. This suggests that guilt induction may be the most effective when an agent knows he/she will suffer over another's failure to act in a desired fashion and is able to explicitly convey this degree of suffering. However, if an agent knows he/she is not going to suffer, then not conveying that information may be less effective.

5 Conclusion

The main motivations of this study were to (i) experimentally test whether agents attempted to exploit the guilt aversion of others in self-serving ways by inducing guilt upon them, (ii) whether agents where susceptible to this type of exploitation, and (iii) whether having such an opportunity fostered more trusting behavior. The experimental data is consistent with Player Bs being susceptible to guilt induction by Player As. This susceptibility of Player Bs to guilt induction is reinforced by data from the post-decision questionnaire where Player Bs' self-reported feelings of guilt where consistent with guilt aversion. However, the data reveals that Player As did not attempt to fully exploit the guilt aversion of Player Bs in the setting considered, despite the fact that it would have increased the likelihood that Player B would have choosen the kind action. Although, the data from the post-decision questionnaire reveals that there may be two *types* of Player As: (i) Player As who think inducing guilt would be effective and, thus, attempt to do it, and (ii) Player As who don't think inducing guilt would be effective and, therefore, do not attempt to do it. Lastly, the data reveals evidence that Player As may be marginally more trusting when playing a trust game where the strategic structure is rich enough for the opportunity to induce guilt upon Player Bs.

The susceptibility of Player Bs to guilt induction that is observed in the data can be viewed as additional experimental evidence consistent with the hypothesis that agents are motivated by guilt aversion. Hence, the experimental design provides an alternative approach for investigating guilt aversion from the previously implemented belief elicitation based approaches, both of which present previously established limitations.[27] The ability to test for guilt aversion without eliciting beliefs is particularly relevant in light of the recent studies by Reuben et al. (2009) and Ellingsen et al.

[27] Specifically, Dufwenberg and Gneezy (2000), Charness and Dufwenberg (2006), Bacharach et al. (2007), and Dufwenberg et al. (2011) elicited second order beliefs and test for a positive correlation between elicited second order expectations and actions. However, because these studies provide only a correlation between elicited second order beliefs and actions, anchoring and false consensus effects cannot be ruled out as possible explanations. Alternatively, Reuben et al. (2009) and Ellingsen et al. (2010) elicit first-order expectations of subjects, convey those expectations to the subject's partner, and test for correlations between expectations and actions. However, the possibility of untruthful reporting of beliefs and skepticism of conveyed beliefs arise with this approach, as noted by Reuben et al.

(2010), which both test for the presence of guilt aversion using similar experimental designs that feature conveyance of elicited beliefs, yet reach opposing conclusions. This paper joins Nelissen at al. (2011) and Charness and Dufwenberg (2011), in its ability to identify behavior consistent with belief dependent models of utility without having to elicit or convey elicited beliefs.

BSH (1994) note that "guilt [induction] does not depend on formal power or influence and may even work best in the absence of such power, because one induces guilt by depicting oneself as the helpless victim of another's actions" (p. 247). This suggests that guilt induction could be particularly effective in economies with less developed legal systems. In such economies, guilt induction could serve as an informal mechanism for enforcing contracts and mitigating corrupt behavior, which might otherwise transpire in the absence of formal prohibitive legislation (Lee, 2010). Guilt induction could also prove to be effective at influencing behavior and impacting outcomes in credence goods markets (Dulleck and Kerschbamer, 2006; Dulleck et al., 2011; Beck et al., 2013; and Balafoutas et al., 2013). With credence goods (e.g., doctors, mechanics, taxis, or other expert services), the consumer is often the "helpless victim" of the expert's actions. Guilt induction by the consumer could be implemented to thwart opportunistic behavior by the expert, especially in developing economies where the incentives for opportunistic behavior are likely to be much stronger.[28]

Partnerships, principle-agent contracting, and employee-employer relationships represent some of the many economic settings where trust is pivotal for successful and efficient relations. There is a growing body of literature that investigates the importance of trust in social and economic settings, and how trust can be fostered (see Fehr, 2009; Charness et al., 2011 for reviews). Much of this literature focuses on the effectiveness of reputation building in fostering trust.[29] While there is often an incentive to trust in economic settings, this incentive is often offset by exposure to the risk of opportunistic behavior by the trusted agent. However, guilt induction by the trusting agent can serve as a mechanism for thwarting such opportunistic behavior, thus mitigating the risk associated with trusting actions. Therefore, having an opportunity to induce guilt would then lead to more trusting behavior, which is what is observed in this paper. This might help shed light on why trust is so prevalent in many economic interactions in our society today, where the strategic environments are often rich enough to allow the possibility to induce guilt.

In his seminal work, Rabin (1993, p. 1296) raises a concluding question about whether agents can "force" emotions in sequential move games? The answer to this question is of clear importance in determining economic outcomes, given that emotions can impact strategic decision making in very systematic and considerable ways. Along these lines, Gneezy and Imas (2014) provide experimental evidence that agents strategically anger others (i.e., they force anger upon others) in self-serving ways. The results from this study suggest that agents can also force guilt upon others, which can then lead to more kind or favorable behavior by those other agents. While this is suggestive that it may be possible for agents to force or manipulate other emotions in sequential move games in self-serving ways, this remains an interesting and open empirical question for future research.

[28] Evidence of such opportunistic behavior by experts in credence goods has been found in a recent field experiment by Balafoutas et al. (2013). In particular, the authors find evidence of systematic over-charging and over-treating of passengers by taxi drivers in Greece.

[29] That is, building a trustworthy reputation through prior trustworthy actions, that are observable to other agents, induces agents to trust you in the future. Many experimental studies have found evidence consistent with this "indirect reciprocity" including Bohnet and Huck (2004), Bolton et al. (2005), Greiner and Levati (2005), Seinen and Schram (2006), Duffy et al. (2008), Engelmann and Fischbacher (2009), and Huck et al. (2012). Charness et al. (2011) provide a thorough review of much of this literature as well as provide experimental evidence that a reputation of trusting behavior can foster trust.

I conclude by noting that the effectiveness of guilt induction as an influence mechanism in strategic settings may have limitations. In particular, repeated applications of guilt induction may become less effective since the target of the guilt induction will likely become resentful or angered over its repeated application. This could ultimately lead to fewer kind actions in response to guilt induction, which is counter to its intended purpose. BSH (1995) recognize this and argue that "although guilt may often be an effective way of getting one's way, it appears to be costly and to carry some stigma. This suggests that inducing guilt may be a technique that has to be used with caution and restraint" (p. 184). Perhaps guilt induction in strategic economic settings should be a mechanism that is reserved for instances when there is little scope for reputational effects, and when the payoff and potential risk associated with a trusting action are largest.

References

[1] Attanasi, G., Battigalli, P., & Manzoni, E. (2014a). "Incomplete Information Models of Guilt Aversion in the Trust Game." IGIER Working Paper n. 480, Bocconi University.

[2] Attanasi, G., Battigalli, P., & Nagel, R. (2014b). "Disclosure of Belief-Dependent Preferences in a Trust Game." IGIER Working Paper n. 506, Bocconi University.

[3] Bacharach, M., Guerra, G., & Zizzo, D. (2007). "The Self-Fulfilling Property of Trust: An Experimental Study." Theory and Decision 63, 349-388.

[4] Balafoutas, L., Beck, A., Kerschbamer, R., & Sutter, M. (2013). "What Dirves Taxi Drivers? A Field Experiment on Fraud in a Market for Credence Goods." Review of Economic Studies 80, 876-891.

[5] Battigalli, P., & Dufwenberg, M. (2007). "Guilt in Games." American Economic Review 97, 170-176.

[6] Battigalli, P., & Dufwenberg, M. (2009). "Dynamic Psychological Games." Journal of Economic Theory 144, 1-35.

[7] Baumeister, R., Stillwell, A., & Heatherton, T. (1994). "Guilt: An Interpersonal Approach." Psychological Bulletin 115, 243-267.

[8] Baumeister, R., Stillwell, A., & Heatherton, T. (1995). "Personal Narratives About Guilt: Role in Action Control and Interpersonal Relationships." Basic and Applied Social Psychology 17, 173-198.

[9] Beck, A., Kerschbamer, R., Qiu, J., & Sutter, M. (2013). "Shaping Beliefs in Experimental Markets for Expert Services: Guilt Aversion and the Impact of Promises and Money-burning Options." Games and Economic Behavior 81, 145-164.

[10] Bellemare, C., Sebald, A., & Strobel, M. (2011). "Measuring the Willingness to Pay to Avoid Guilt: Estimation using Equilibrium and Stated Belief Models." Journal of Applied Econometrics 26, 437-453.

[11] Bellemare, C., Kröger, S., & Van Soest, A. (2008). "Measuring Inequity Aversion in a Heterogeneous Population using Experimental Decisions and Subjective Probabilities." Econometrica 76, 815-839.

[12] Berg, J., Dickhaut, J. & McCabe, K. (1995). "Trust, Reciprocity and Social History." Games and Economic Behavior 10, 122-142.

[13] Bohnet, I., & Huck, S. (2004). "Repetition and Reputation: Implications for Trust and Trustworthiness when Institutions Change." American Economic Review Papers and Proceedings 94, 362-366.

[14] Bolton, G., Katok, E., & Ockenfels, A. (2005). "Cooperation Among Strangers with Limited Information About Reputation." Journal of Public Economics 89, 1457-1468.

[15] Bolton, G., & Ockenfels, A. (2000). "A Theory of Equity, Reciprocity, and Competition." American Economic Review 90, 166-193.

[16] Brandts, J., & Charness, G. (2003). "Truth or consequences: An experiment." Management Science 49, 116-130.

[17] Charness, G., Du, N., & Yang, C. (2011). "Trust and Trustworthiness Reputations in an Investment Game." Games and Economic Behavior 72, 361-375.

[18] Charness, G., & Dufwenberg, M. (2006). "Promises and Partnership." Econometrica 74, 1579–1601.

[19] Charness, G., & Dufwenberg, M. (2010). "Bare Promises: An Experiment." Economics Letters 107, 281-283.

[20] Charness, G., & Dufwenberg, M. (2011). "Participation." American Economic Review 101, 1211-1237.

[21] Charness, G., & Rabin, M. (2002). "Understanding Social Preferences with Simple Tests." Quarterly Journal of Economics 117, 817-869.

[22] Cox, J. (2004). "How to Identify Trust and Reciprocity." Games and Economic Behavior 46, 260-281.

[23] De Hooge, I. E., Nelissen, R., Breugelmans, S. M., & Zeelenberg, M. (2011). "What is Moral about Guilt? Acting "Prosocially" at the Disadvantage of Others." Journal of Personality and Social Psychology 100, 462.

[24] Duffy, J., Lee, Y., & Xie, H. (2008). "Social Norms, Information, and Trust Among Strangers: An Experimental Study." Mimeo.

[25] Dufwenberg, M. (2002). "Marital Investments, Time Consistency, and Emotions." Journal of Economic Behavior and Organization 48, 57-69.

[26] Dufwenberg, M., Gächter, S., & Hennig-Schmidt, H. (2011). "The Framing of Games and the Psychology of Play." Games and Economic Behavior (In Press).

[27] Dufwenberg, M. & Gneezy, U. (2000). "Measuring beliefs in an experimental lost wallet game." Games and Economic Behavior 30, 163-182.

[28] Dufwenberg, M. & Kirchsteiger, G. (2004). "A Theory of Sequential Reciprocity." Games and Economic Behavior 47, 268-298.

[29] Dufwenberg, M., Smith, A., & Van Essen, M. (2013). "Hold-Up: With a Vengeance." Economic Inquiry 51, 896-908.

[30] Dulleck, U. & Kerschbamer, R. (2006). "On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods." Journal of Economic Literature 44, 5-42.

[31] Dulleck, U., Kerschbamer, R., & Sutter, M. (2011). "The Economics of Credence Goods: On the Role of Liability, Verifiability, Reputation and Competition." American Economic Review 101, 526-555.

[32] Ellingsen, T., Johannesson, M., Tjotta, S., & Torsvik, G. (2010). "Testing Guilt Aversion." Games and Economic Behavior, 68, 95-107.

[33] Engelmann, D., & Fischbacher, U. (2009). “Indirect Reciprocity and Strategic Reputation Building in an Experimental Helping Game.” Games and Economic Behavior 67, 399-407.

[34] Engelmann, D., & Strobel, M. (2004). “Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments.” American Economic Review 94, 857-869.

[35] Erat, S., & Gneezy, U. (2012). “White Lies.” Management Science 58, 723-733.

[36] Fehr, E. (2009). “On the Economics and Biology of Trust.” Journal of the European Economic Association 7, 235-266.

[37] Fehr, E., & Schmidt, K. (1999). “A Theory of Fairness, Competition, and Cooperation.” Quarterly Journal of Economics 114, 817-868.

[38] Fischbacher, U. (2007). “z-Tree, Toolbox for Readymade Economic Experiments.” Experimental Economics 10, 171–178.

[39] Fong, Y., Huang, C., & Offerman, T. (2007). “Guilt Driven Reciprocity in a Psychological Signaling Game.” Working Paper.

[40] Geanakopolos, J., Pearce, D. & Stacchetti, E. (1989). “Psychological Games and Sequential Rationality.” Games and Economic Behavior 1, 60-79.

[41] Gneezy, U. (2005). “Deception: The Role of Consequences.” American Economic Review 95, 384-394.

[42] Gneezy, U., & Imas, A. (2014). “Materazzi Effect and the Strategic use of Anger in Competitive Interactions.” Proceedings of the National Academy of Sciences 111.4, 1334-1337.

[43] Greiner, B., & Levati, V. (2005). “Indirect Reciprocity in Cyclical Networks: An Experimental Study.” Journal of Economic Psychology 26, 711-731.

[44] Güth, W., Mugera, H., Musau, A., & Ploner, M. (2013). “Deterministic versus Probabilistic Consequences of Trust and Trustworthiness: An Experimental Investigation.” Journal of Economic Psychology, In Press.

[45] Hoffman, M. L. (1982). Development of prosocial motivation: Empathy and guilt. In N. Eisenberg (Ed.), The development of prosocial behavior (pp. 281-313). New York: Academic Press.

[46] Huck, S., Lünser, G., & Tyran, J-R. (2012). “Competition Fosters Trust.” Games and Economic Behavior 76, 195-209.

[47] Lee, S.Y., (2010). “Economics of Guanxi as an Interpersonal Investment Game.” Review of Development Economics 14, 333-342.

[48] Nelissen, R., Leliveld, M. C., van Dijk, E., & Zeelenberg, M. (2011). “Fear and Guilt in Proposers: Using Emotions to Explain Offers in Ultimatum Bargaining.” European Journal of Social Psychology 41, 78-85.

[49] Rabin, M. (1993). “Incorporating Fairness into Game Theory and Economics.” American Economic Review 83, 1281-1302.

[50] Reuben, E., Sapienza, P., & Zingales, L. (2009). “Is Mistrust Self-Fulfilling?” Economic Letters 104, 89-91.

[51] Seinen, I., & Schram, A. (2006). "Social Status and Group Norms: Indirect Reciprocity in a Helping Experiment." European Economic Review 50, 581-602.

[52] Tangney, J. (1990). "Assessing Individual Differences in Proneness to Shame and Guilt: Development of the Self-Conscious Affect and Attribution Inventory." Journal of Personality and Social Psychology 59, 102-111.

[53] Tangney, J., & Dearing, R. (2002). Shame and guilt. New York, NY: Guilford Press.

[54] Tangney, J., & Fischer, K. (1995). Self-conscious emotions: Shame, guilt, embarrassment, and pride. New York: Guilford Press.

[55] Vangelisti, A., Daly, J., & Rudnick, J. (1991). "Making People Feel Guilty in Conversations: Techniques and Correlates." Human Communication Research 18, 3-39.

[56] Wang, C. S., Galinsky, A. D., & Murnighan, J. K. (2009). "Bad Drives Psychological Reactions, but Good Propels Behavior Responses to Honesty and Deception." Psychological Science 20, 634-644.

[57] Wang, C. S., & Leung, A. K. Y. (2010). "The Cultural Dynamics of Rewarding Honesty and Punishing Deception." Personality and Social Psychology Bulletin 36, 1529-1542.

Appendix A – Application of B&D Model of Guilt

In this Appendix, I formalize the arguments put forth in Section 2.3 regarding the connection between the posited method by BSH (1994) of how guilt is induced and the formal model of guilt developed by B&D. In doing so, I first provide a brief outline of the B&D model of guilt, followed by its application to Γ_{PPT}. I then derive conditions under which the BSH (1994) method is consistent with predictions of the B&D model of guilt. I conclude by showing that the BSH (1994) method can be supported as an equilibrium in Γ_{PPT} under the framework of the B&D model. I conclude by briefly discussing the recent extention of the B&D model developed by Attanasi et a. (2014a) that incorporates incompelete information about guilt sensitivities, and how incomplete infomation about guilt sensitivies can impact behavioral predicitions in Γ_{PPT}.

B&D Model of Guilt Applied to Γ_{PPT}:

Before I proceed in formally applying the B&D model of guilt to Γ_{PPT}, I first provide a general overview of the B&D model of simple guilt.[30] What follows is only a simplified outline of the model. Interested readers should refer to B&D for the full, technical presentation of the model including illustrative examples. Informally, the model posits that agents suffer disutility, in the form of guilt, from failing to live up to others' expectations. This is captured by modeling an agent's utility as a function of his/her own material payoffs and the extent to which he/she let other agents down.

Formally, simple guilt is modeled by specifying a utility function for player i given by:

$$u_i^{SG} = m_i - \sum_{j \neq i} \theta_{ij} \cdot D_j \qquad \text{(Simple Guilt Utility)}$$

In this expression, m_i represents player $i's$ material payoff, and $\sum\limits_{j \neq i} \theta_{ij} \cdot D_j$ represents player $i's$ disutility from simple guilt. The latter component is composed of two pieces. The first, θ_{ij}, is an exogenously given constant that measures player $i's$ sensitivity of feeling guilty toward player j. The second, D_j, represents the amount by which player i lets player j down, as a result of player $i's$ strategy. $D_j = E_j - m_j$ is expressed as the difference between the material payoff that player j was expecting, E_j, and the material payoff that player j actually receives, m_j. E_j itself is a function of player $j's$ strategy, and player $j\prime s$ vector of "first-order" beliefs regarding the strategies of the other players. Note, player i does not actually observe D_j, as it is a function of player $j's$ first-order beliefs. Therefore, it is assumed that player i maximizes the expected value of u_i^{SG}, given player $i's$ first-order beliefs regarding player $j's$ strategy, and player $i's$ "second-order" belief regarding player $i's$ first-order beliefs.

In proceeding with the application of the B&D model of simple guilt to Γ_{PPT}, I derive the guilt that Player B would experience from choosing *Left* in Γ_{PPT} at each of the three possible conveyance states (histories): (i) $X = 0$ was conveyed, which I denote C^0, (ii) $X = 6$ was conveyed, which I denote C^6, and (iii) the value of X was not conveyed, which I denote C^N. A strategy for Player B is a probability distribution over Player B's possible actions, $\{$*Left, Right*$\}$, at each of the three

[30] B&D model two types of guilt for a general class of extensive form games, simple guilt and guilt from blame. With simple guilt, an agent suffers disutility proportional to how much he/she lets down another agent. However, with guilt from blame, an agent suffers disutility proportional to how much the other agent blames him/her for being let down. Thus, the main difference between the two models is the extent to which an agent can be blamed for letting down another agent. With respect to Γ_{PPT}, the two models are equivalent. Player A can unambiguously identify the action of Player B, which implies that Player B will receive all the blame for letting Player A down. Although either of these models of guilt could be used for this analysis, I opt to apply the less complex model of simple guilt for clarity.

possible conveyance states. In deriving this guilt that Player B would suffer from choosing *Left*, it is necessary to first derive Player A's material expectation. Before Player A makes her initial *In* or *Out* decision, Player A forms an initial first-order belief regarding Player B's strategy, which can be represented as the probabilities that Player B would choose *Right* at each of the three conveyance states; I denote this vector of probabilities as $\boldsymbol{\alpha}_A$= (Pr($Right|C^0$),Pr($Right|C^N$),Pr($Right|C^6$)). For a given strategy, Player A forms an initial expectation, which I denote E_A, weighted over $\boldsymbol{\alpha}_A$ and moves by Nature, of her material payoff.

Player B experiences disutility from simple guilt when he chooses a strategy that yields a payoff to Player A that is lower than E_A. However, Player B does not observe E_A. Therefore, Player B must form an expectation of E_A, conditional on the conveyance state. I define this conditional expectation of E_A as $E_B|h \textit{ where } h \in \{C^0, C^N, C^6\}$, which is implicitly a function of Player B's conditional second-order beliefs regarding $\boldsymbol{\alpha}_A$. The guilt that Player B would suffer from choosing *Left* is proportional to the difference between $E_B|h$ and m_A, where m_A is the material payoff that Player A actually receives as a result of Player B's *Left* decision. Let $\theta_B \geq 0$ denote Player B's sensitivity to feeling guilty. Below is the amout of simple guilt that Player B would suffer from choosing *Left* at each of the three possible conveyance states.

X = 0 was conveyed: Because the conveyance is credible, Player B knows that if he chooses *Left,* Player A will receive a payoff of $m_A = 0$. Player B's expectation of Player A's expectation is $E_B|C^0$. By choosing *Left,* Player B will suffer disutility from guilt equal to $\theta_B \cdot (E_B|C^0 - 0)$. Thus, Player B's utility from choosing *Left*, after $X = 0$ was conveyed, is equal to:

$$6 - \theta_B \cdot (E_B|C^0 - 0)$$

X = 6 was conveyed: Because the conveyance is credible, Player B knows that if he chooses *Left,* Player A will receive a payoff of $m_A = 6$. Player B's expectation of Player A's expectation is $E_B|C^6$. By choosing *Left,* Player B will suffer disutility from guilt equal to $\theta_B \cdot (E_B|C^6 - 6)$. Thus, Player B's utility from choosing *Left*, after $X = 0$ was conveyed, is equal to:

$$6 - \theta_B \cdot (E_B|C^6 - 6)$$

Value of X not conveyed If the value of X is not conveyed, then Player B must think about the expected material payoff that Player A would receive if he chooses *Left.* Let $\widehat{m}_A = E_B[m_A|C^N]$ denote this expectation. As I have previously shown (see Section 2.1), $\widehat{m}_A \in [1, 5]$. Player B's expectation of Player A's expectation is $E_B|C^N$. By choosing *Left*, Player B will suffer disutility from guilt equal to: $\theta \cdot (E_B|C^N - \widehat{m}_A)$ where $\widehat{m}_A \in [1, 5]$. Thus, Player B's utility from choosing *Left*, after the value of X was not conveyed, is equal to:

$$6 - \theta_B \cdot (E_B|C^N - \widehat{m}_A) \textit{ where } \widehat{m}_A \in [1, 5]$$

Guilt Induction in Γ_{PPT} via the B&D Model of Guilt:

Based on the BSH (1994) method for how an agent can iduce guilt upon another, it was derived that Player A could attempt to induce guilt upon Player B in Γ_{PPT} by choosing to *Convey* $X = 0$ and *Not Convey* $X = 6$. In what follows, I derive conditions under which this is consistent with predictions of the B&D model of simple guilt. From the previous section, we derived that the disutilities, from feeling guilt, that Player B would suffer from choosing *Left* at each of the three possible conveyance states are:

- $X = 0$ was conveyed: $\theta_B \cdot (E_B|C^0 - 0)$
- $X = 6$ was conveyed: $\theta_B \cdot (E_B|C^6 - 6)$
- Value of X not conveyed: $\theta_B \cdot (E_B|C^N - \widehat{m}_A)$ where $\widehat{m}_A \in [1, 5]$

According to B&D, Player B would suffer more disutility from choosing *Left* when $X = 0$ was conveyed, compared to when the value of X was not conveyed, when:

$$\theta_B \cdot (E_B|C^0 - 0) \geq \theta_B \cdot (E_B|C^N - \widehat{m}_A) \Longrightarrow E_B|C^0 \geq E_B|C^N - \widehat{m}_A \qquad \text{(Condition 1)}$$

Similarly, Player B would suffer more disutility from choosing *Left* when the value of X was not conveyed, compared to when $X = 6$ was conveyed, when:

$$\theta_B \cdot (E_B|C^N - \widehat{m}_A) \geq \theta_B \cdot (E_B|C^6 - 6) \Longrightarrow E_B|C^N - \widehat{m}_A \geq E_B|C^6 - 6 \qquad \text{(Condition 2)}$$

Essentially, Condition 1 states that Player B's second-order belief of Player A's expectation after $X = 0$ is conveyed, $E_B|C^0$, is not *too* much lower than Player B's second-order belief of Player A's expectation after the value of X is not conveyed, $E_B|C^N$. Similarly, Condition 2 states that Player B's second-order belief of Player A's expectation after the value of X is not conveyed, $E_B|C^N$, is not *too* much lower than Player B's second-order belief of Player A's expectation after $X = 6$ is conveyed, $E_B|C^6$. Conditions 1 and 2 would certainly be satisfied if we assumed that Player B does not update his belief of Player A's expectation (i.e., $E_B|C^0 = E_B|C^N = E_B|C^6$), which would be satisfied in an equilibrium. However, making the assumption that $E_B|C^0 = E_B|C^N = E_B|C^6$ is clearly stronger than is needed for Condition 1 and 2 to be satisfied.

If Conditions 1 and 2 are satisfied, then the B&D model predicts that Player B would feel more guilt from choosing left after $X = 0$ was conveyed and the value of X was not conveyed, compared to when the value of X was not conveyed and $X = 6$ was conveyed, respectively. Therefore, if Condition 1 and 2 hold, then according to the B&D model, Player A can induce guilt upon Player B by choosing to *Convey* $X = 0$ and *Not Convey* $X = 6$, which is consistent with the BSH (1994) method for inducing guilt.

Guilt Induction as an Equilibrium of Γ_{PPT} :

Next, I show that effective guilt induction in Γ_{PPT} can be supported as sequential equilibrium (SE) of Γ_{PPT} under the framework of B&D.[31] The intuition behind this rests in the fact that in a SE, an assessment (profile of behavioral strategies and conditional hierarchical beliefs) will be consistent. Battigalli and Dufwenberg (2009) show that in equilibrium, "players never change their beliefs about the conditional beliefs that the opponents would hold at each h (history)" (pp. 16). Thus, in equilibrium, there is no belief updating. It follows that $E_B|C^0 = E_B|C^N = E_B|C^6$, which implies that Conditions 1 and 2 from above would be satisfied in an equilibrium. In this equilibrium analysis, I will assume that there is complete information regarding Player B's guilt sensitity, θ_B. While this assumption may be unrealistic, and not a feature of the design, it is common in the extant literature related to belief dependent motivations (e.g., Charness and Dufwenberg, 2006; 2011; Dufwenberg et al., 2013); furthermore, it allows me to focus on the motivation of this analysis, which is to establish that guilt induction can be supported as an equilibrium, without the

[31] I refer the reader to the authors' more general paper, Battigalli and Dufwenberg (2009), for a formal equilibrium analysis of *dynamic psychological games*. The authors extend the concept of sequential equilibrium by incorporating hierarchies of conditional beliefs.

information complexities that would otherwise need to be addressed. I aslo assume that Player A is not guilt averse, i.e., Player A's guilt sensitivity is zero.[32]

Guilt induction by Player A in Γ_{PPT} is characterized by the choice to *Convey* $X = 0$ and *Not Covey* $X = 6$. Therefore, the strategy (*In, Convey* $X = 0$, *Not Convey* $X = 6$) for Player A is consistent with attempted guilt induction. Effective guilt induction is characterized by Player B choosing *Right* as a response to the guilt induction by Player A. The following two strategies for Player B are consistent with responding in kind to Player A's guilt induction: (*Right*$|C^0$, *Right*$|C^N$, *Left*$|C^6$), and (*Right*$|C^0$, *Left*$|C^N$, *Left*$|C^6$). Thus, the strategy profiles ((*In, Convey* $X = 0$, *Not Convey* $X = 6$), (*Right*$|C^0$, *Right*$|C^N$, *Left*$|C^6$)), and ((*In, Convey* $X = 0$, *Not Convey* $X = 6$), (*Right*$|C^0$, *Left*$|C^N$, *Left*$|C^6$)) are the candidate equilibrium profiles for effective guilt induction in Γ_{PPT}. In Claims 1 and 2 below, I show that each of these strategy profiles can be supported as an equilibrium of Γ_{PPT}.

Claim 1 The strategy profile ((*In, Convey* $X = 0$, *Not Convey* $X = 6$), (*Right*$|C^0$, *Right*$|C^N$, *Left*$|C^6$)) can be supported as a SE of Γ_{PPT} for $\theta_B \in [\frac{2}{5}, \frac{1}{2}]$

To verify that this strategy profile is an equilibrium, we need to check that neither player has a profitable deviation. For Player A, this is rather trivial. By following the equilibrium strategy, Player A earns a payoff of 10, which is the highest payoff of the game. Therefore, Player A has no profitable deviation. For Player B, we need to consider deviations at each of the possible conveyance states. Given consistent beliefs in equilibrium, we have that $\boldsymbol{\alpha}_A = \boldsymbol{\beta}_B = (1, 1, 0)$, $E_A = E_B|h = 10 \ \forall \ h \in \{C^0, C^N, C^6\}$, and $\widehat{m}_A = 5$. Player B will not deviate to *Right*$|C^6$ so long as: $6 - \theta_B \cdot [10 - 6] \geq 4 \Leftrightarrow \theta_B \leq \frac{1}{2}$. Player B will not deviate to *Left*$|C^N$ so long as: $4 \geq 6 - \theta_B \cdot [10 - 5] \Leftrightarrow \theta_B \geq \frac{2}{5}$. Similarly, Player B will not deviate to *Left*$|C^0$ so long as: $4 \geq 6 - \theta_B \cdot [10 - 0] \Leftrightarrow \theta_B \geq \frac{1}{5}$ which is satisfied if $\theta_B \geq \frac{2}{5}$.

Claim 2 The strategy profile ((*In, Convey* $X = 0$, *Not Convey* $X = 6$), (*Right*$|C^0$, *Left*$|C^N$, *Left*$|C^6$)) can be supported as a SE of Γ_{PPT} for $\theta_B \in [\frac{2}{7}, 1]$

Again, to verify that this strategy profile is an equilibrium, we need to check that neither player has a profitable deviation. For Player A, playing the equilibrium strategy yields an expected payoff of 7; therefore, Player A cannot profitably deviate to *Out.* Player A would not deviate and *Not Convey* $X = 0$, which would result in a payoff of 0 compared to a payoff of 10 from following the equilibrium strategy to *Convey* $X = 0$. Player A is indifferent between *Convey* $X = 6$ and *Not Convey* $X = 6$. Therefore, Player A has no profitable deviation from the prescribed equilibrium strategy. For Player B, we need to consider deviations at each of the possible conveyance states. Given consistent beliefs in equilibrium, we have that $\boldsymbol{\alpha}_A = \boldsymbol{\beta}_B = (1, 0, 0)$, $E_A = E_B|h = 7 \ \forall \ h \in \{C^0, C^N, C^6\}$, and $\widehat{m}_A = 5$. Player B will not deviate to *Right*$|C^6$ so long as: $6 - \theta_B \cdot [7 - 6] \geq 4 \Leftrightarrow \theta_B \leq 2$. Player B will not deviate to *Right*$|C^N$ so long as: $6 - \theta_B \cdot [7 - 5] \geq 4 \Leftrightarrow \theta_B \leq 1$. Similarly, Player B will not deviate to *Left*$|C^0$ so long as: $4 \geq 6 - \theta_B \cdot [7 - 0] \Leftrightarrow \theta_B \geq \frac{2}{7}$.

[32] This assumption that Player A (the trustor) does not experience guilt is consistent with the model of "role-dependent guilt" developed by Attanasi et al. (2014a). Because participants are randomly drawn from the same population to play the role of Player A or Player B, this assumption is consistent with the idea that the guilt sensitivity of a participant becomes an actual tendency when they are assigned the role of Player B, as asserted by Attanasi et al. (2014a). This assumption that the trustor is not affected by guilt, which is common knowledge, is also employed by Attanasi et al. (2014b).

Incomplete Information Regarding Sensitivity to Guilt:

As I stated in the previous section, equilibrium Claims 1 and 2 assume common knowledge of Player B's sensitivity to guilt, θ_B. In a recent paper, Attanasi et al. (2014a) have extended the B&D model to allow for incomplete information regarding guilt sensitivities in a 2-player trust game.[33] A formal application of their model to Γ_{PPT} would require extensive theoretical preliminaries regarding the formulation of hierarchies of beliefs for each player via "type" structures, which is beyond the scope of this paper and would not add a proportionate increase in the main conclusions drawn. In addition, because I do not elicit any information about beliefs or types in the experiment (as the sessions were run prior the Attanasi et al., 2014a paper), I would be unable to derive precise predictions of the incomplete information model. As a result, forgo a formal equilibrium analysis of Γ_{PPT} under incomplete information; rather, I appeal to insights from the Attanasi et al. (2014a) paper to provide some general discussion regarding how incomplete information about types could impact behavior in Γ_{PPT}.[34]

The main difference between the complete information and incomplete information extension of the guilt aversion model (as is most relevant to the discussion that follows) is that in an incomplete information setting where Player As and Player Bs are matched at random (as in the current experimental procedure), it is likely that participants hold heterogeneous and dispersed beliefs (Attanasi et al., 2014b p. 18). The important behavioral implication is that under complete information, there is likely to be much more polarization of behavior because the common knowledge of θ_B enables coordination. Specifically, if θ_B is *small enough* then Player B will not be susceptible to guilt induction and will choose *Left,* and Player A will choose *Out* (the standard payoff maximizing equilibrium outcome)*;* on the other hand, if θ_B is *sufficiently large* then Player B will be susceptible to guilt induction by Player A, and Player A could increase his/her payoff by choosing *In* and then inducing guilt (consistent with Claims 1 and 2 above).

However, with incomplete information about θ_B (a reasonable assumption with regard to Player As in the experiment) more dispersion in beliefs and mis-coordination can arise, which can possibly lead to interesting patterns of behavior in Γ_{PPT}. Namely, many Player As could have held beliefs that Player Bs were not sensitive to guilt (i.e., low θ_B), which may have motivated them to choose *Out,* which could explain why there was only a marginal 14% increase in *In* rates between Γ_{PPT} and Γ_{UPT}. Additionally, for those Player As that did choose *In* (for other possible reasons besides attempting to exploit the guilt aversion of Player B), they would not be motivated to convey $X = 0$ and/or not convey $X = 6$ as a means of inducing guilt, as it would not be effective given their beliefs about θ_B (which is consistent with the analysis of the questionnaire data in Table 4); these Player As may opt, for example, to *Convey* $X = 6$ to display honesty to Player B, consistent with the discussion in Section 4.3. On the other extreme, some Player As could have believed that Player Bs were *very* sensitive to feeling guilt (i.e., high θ_B), such that Player B would be motivated to choose *Right* regardless of the conveyance state; in this case, these Player As would, again, not necessarily be motivated to convey $X = 0$ and/or not convey $X = 6$ as a means of inducing guilt, as it would not be necessary to motivate Player B to choose *Right.*

If we assume that Player As have incomplete information about Player B's guilt sensitivity, θ_B,then multiple equilibria can arise in Γ_{PPT}, as in the case with perfect information. In particular, the "selfish" equilibrium exists that results in Player A always choosing *Out* since Player B would

[33] I thank an anonymous reviewer for calling my attention to this paper, as well as making several apt suggestions of how to incorporate its main insights into the theoretical guilt aversion analysis of Γ_{PPT}.

[34] I refer interested readers to Attanasi et al. (2014a) for the formal development of the incomplete information model of guilt aversion and the corresponding exhaustive equilibrium analysis.

always choose *Left,* if Player Bs have sufficiently low θ_Bs.[35] In addition, by placing some restrictions on the distribution of θ_B (e.g., assuming $\theta_B \in [\frac{2}{7}, 1]$), equilibria that correspond to effective guilt induction (those specified in Claims 1 and 2 above) can also arise. However, with incomplete information regarding θ_B, players do not have the ability to coordinate based on θ_B; hence, it is much more likely to observe heterogeneity in behavior because of the dispersion in beliefs (as shown by Attanasi et al., 2014b). As a result, the fact that participants playing the role of Player A posses incomplete information regarding Player B's sensitivity to guilt could explain some of the observed patterns in the data; namely, why only marginally more Player As chose *In* in Γ_{PPT}, as well as the conveyance behavior inconsistent with attempted guilt induction, where Player As choose to *Not Convey* $X = 0$ and *Convey* $X = 6$.

[35] The condition that Player Bs need to have sufficiently low θ_Bs to ensure the "selfish" equilibrium outcome is a result of the fact that Γ_{PPT} has a chance move. If there were no chance moves, then we would have existence of the selfish equilibrium for all types of Player As and Player Bs (see, specifically, Remark 1 in Attanasi et al. 2014a and Observation 2 from B&D more generally). However, because of the chance move in Γ_{PPT} it is possible for Player A to still be let down from B's choice of *Left* (if chance were to chose $X = 0$), even if Player B believes that Player A expected him/her to choose *Left.* For the specific parameterization of Γ_{PPT}, assuming all Player Bs have $\theta_B \leq \frac{2}{3}$ would be sufficient for the existence of the selfish equilibrium for all types.

Appendix B.1 – Copy of Experimental Instructions

Sample Instructions – PPT Treatment

Welcome and thank you for participating. Your participation is **VOLUNTARY,** and you may leave at any time. Feel free to raise your hand and ask questions at any time, and you may refer back to these instructions at any time during the session. Please remain seated and quiet for the remainder of the session. All decisions are to be completed individually and interaction with other participants is strictly **PROHIBITED.** Thank you for your cooperation.

Each person will receive a $5 show-up payment for participating. In addition, you can receive additional compensation based on the decision(s) that are made in the decision task described below. After the task is complete, you will be privately paid the amount of money you have earned. Upon completions of the decision task, please remain quietly seated in your carrel until you have been paid.

The Decision Task:

You will be participating in a 2-person decision task. Each person will be randomly and anonymously paired with another person in the lab. In each of the 2-person decision-making pairs, one person will be randomly assigned the role of PLAYER A and the other person will be randomly assigned the role of PLAYER B. You will remain in your assigned role for the entire session. The earnings of each Player will depend on the decision(s) he/she makes, and/or the decision(s) of the Player with whom they are paired. A brief outline of steps of the decision task will first be provided, followed by a detailed description of each step and the corresponding earnings for Player.

Step 1: PLAYER A begins by first choosing IN or OUT

- If PLAYER A chooses OUT, the task ends
- If PLAYER A chooses IN, then the task proceeds to Step 2

Step 2: PLAYER A *might* privately learn some payoff information that was initially unknown to both players. If PLAYER A does learn the information, the PLAYER A will then have an opportunity to convey the information to PLAYER B. Then the task will proceed to Step 3.
Step 3: PLAYER B chooses between RIGHT or LEFT, and the task ends

A full description of each step and the corresponding earnings for Player follows:

(1) PLAYER A first chooses between IN or OUT.

- If PLAYER A chooses OUT, then the decision task ends. PLAYER A will receive $6 and PLAYER B will receive $2.
- If PLAYER A chooses IN, the task proceeds to step (2) where PLAYER A might privately learn the unknown information, and then have an opportunity to convey that information to PLAYER B. After step (2), the task will proceed to step (3) where PLAYER B will then be asked to decide between RIGHT or LEFT.

(2) I postpone the details about the information that PLAYER A can possible learn, and convey to PLAYER B until after step (3) is described. Describing step (3) first will help clarify step (2).

(3) If PLAYER A chooses IN, at step (1), then PLAYER B must choose between RIGHT or LEFT.

- If PLAYER B chooses RIGHT, then the decision task ends. PLAYER A will receive $10 and PLAYER B will receive $4.
- If PLAYER B chooses LEFT, then the decision task ends and PLAYER A will receive $X and PLAYER B will receive $6. There is a 50% chance that X = $0 and a 50% chance that X = $6. That is, X = $0 and X = $6 are equally likely.

NOTE: When the decision task begins, neither PLAYER A nor PLAYER B knows the value of X. Therefore, PLAYER A does not know the value of X when he/she decides between IN or OUT in step (1).

Back to the description of step (2):

(2) If PLAYER A chooses IN, there is an **80%** chance that PLAYER A **will** privately learn the value of X, and a **20%** chance that PLAYER A **will not** privately learn the value of X.

- If PLAYER A **does** learn the value of X (80% chance), then PLAYER A must then decide whether or not to convey the value of X to PLAYER B **before** PLAYER B makes his/her decision in step (3).
  - If PLAYER A **does convey** the value of X, then PLAYER B **will know** the value of X before he/she decides between RIGHT or LEFT in step (3).
  - If PLAYER A **does not convey** the value of X, then PLAYER B **will not know** the value of X before he/she decides between RIGHT or LEFT in step (3).

- If PLAYER A **does not** learn the value of X (20% chance), then PLAYER A will not have an opportunity to convey the value of X to PLAYER B. The task will proceed to step (3) where PLAYER B will then choose between RIGHT or LEFT without knowing the value of X.

If PLAYER A chose to convey the value of X, then when PLAYER B makes his/her decision in step (3), the following message will appear on PLAYER B's screen:

"PLAYER A has chosen to convey the value of X to you"
"The value of X is: [actual value of X]"

If either (a) PLAYER A did not learn the value of X after choosing IN, or (b) PLAYER A did learn the value of X but chose not to convey the value of X, then when it is time for PLAYER B to make his/her decision in step (3), the following message will appear on PLAYER B's screen:

"The value of X remains unknown"

Payoff Table:

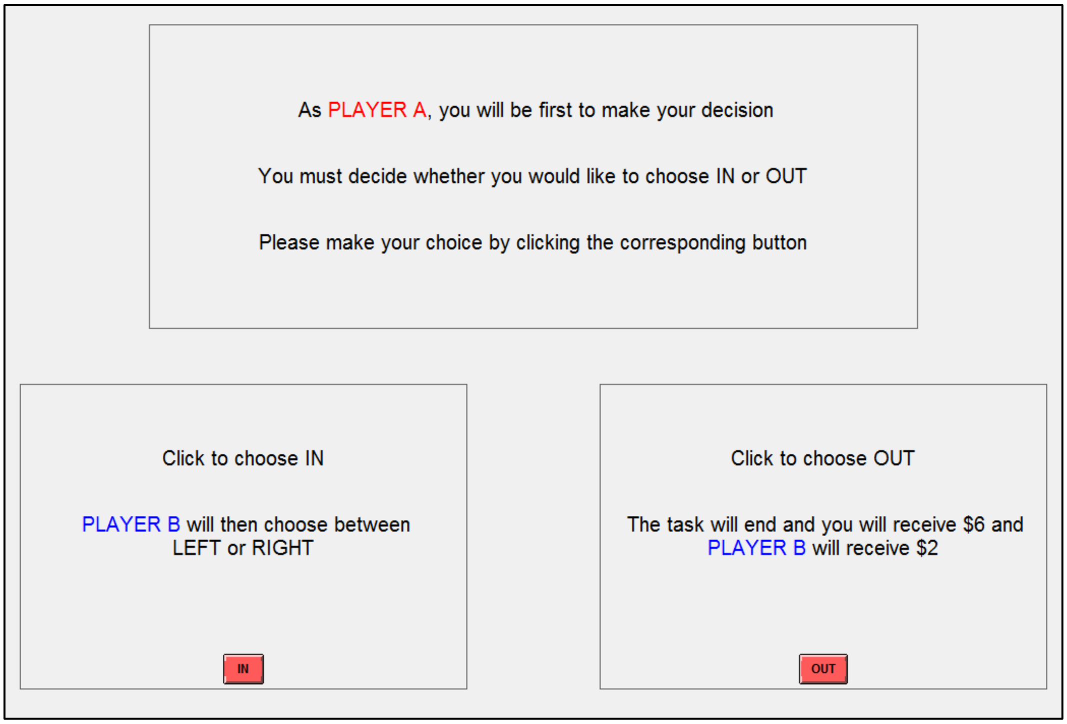
The table below summarizes the earnings of each Player for each of the possible outcomes in the decision task:

DECISION OUTCOME	Earnings of PLAYER A	Earnings of PLAYER B
PLAYER A chooses OUT	$6	$2
PLAYER A first chooses IN and then:		
PLAYER B chooses RIGHT	$10	$4
PLAYER B chooses LEFT	$X	$6
There is a 50% chance X = 0 and a 50% chance X = 6		

Each person will participate in this decision making task ONE time. After the task has ended, the decision(s) of each Player and the corresponding earnings of each Player will be revealed to both Players. Additionally, the value of X will be revealed to both PLAYER A and PLAYER B regardless of the decisions made in the task. You will then be asked to fill out a short questionnaire that will take about 3 minutes to complete. Your answers to the questionnaire are confidential and will not be shared with any other participants. After completion of the questionnaire, an Experimenter will then come by and privately pay you your total experimental earning which equals your earnings from the decision task **PLUS** the $5 show-up payment. After you have been paid, you may quietly exit the lab.

Appendix B.2 – Sample Screen Shots

Player A's Initial *IN/OUT* Decision



As PLAYER A, you will be first to make your decision

You must decide whether you would like to choose IN or OUT

Please make your choice by clicking the corresponding button

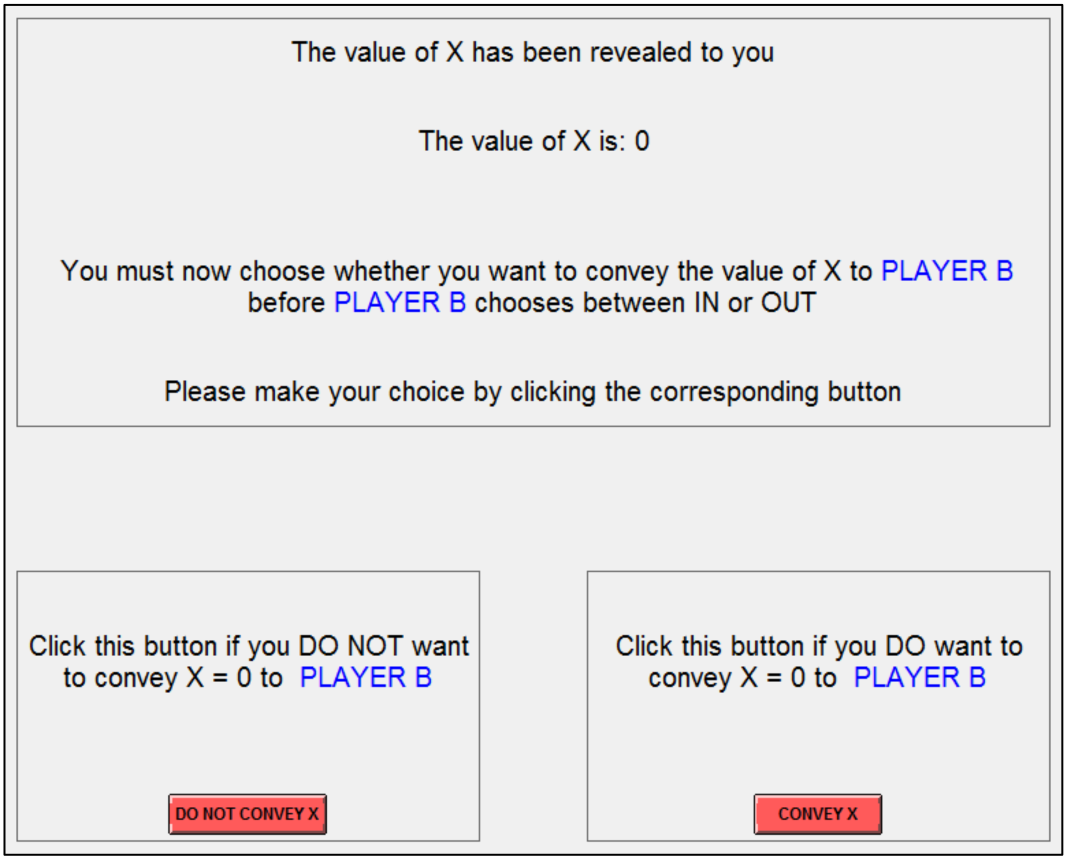
Click to choose IN

PLAYER B will then choose between LEFT or RIGHT

Click to choose OUT

The task will end and you will receive $6 and PLAYER B will receive $2

Player A's Conditional *Convey/Not Convey* Decision



The value of X has been revealed to you

The value of X is: 0

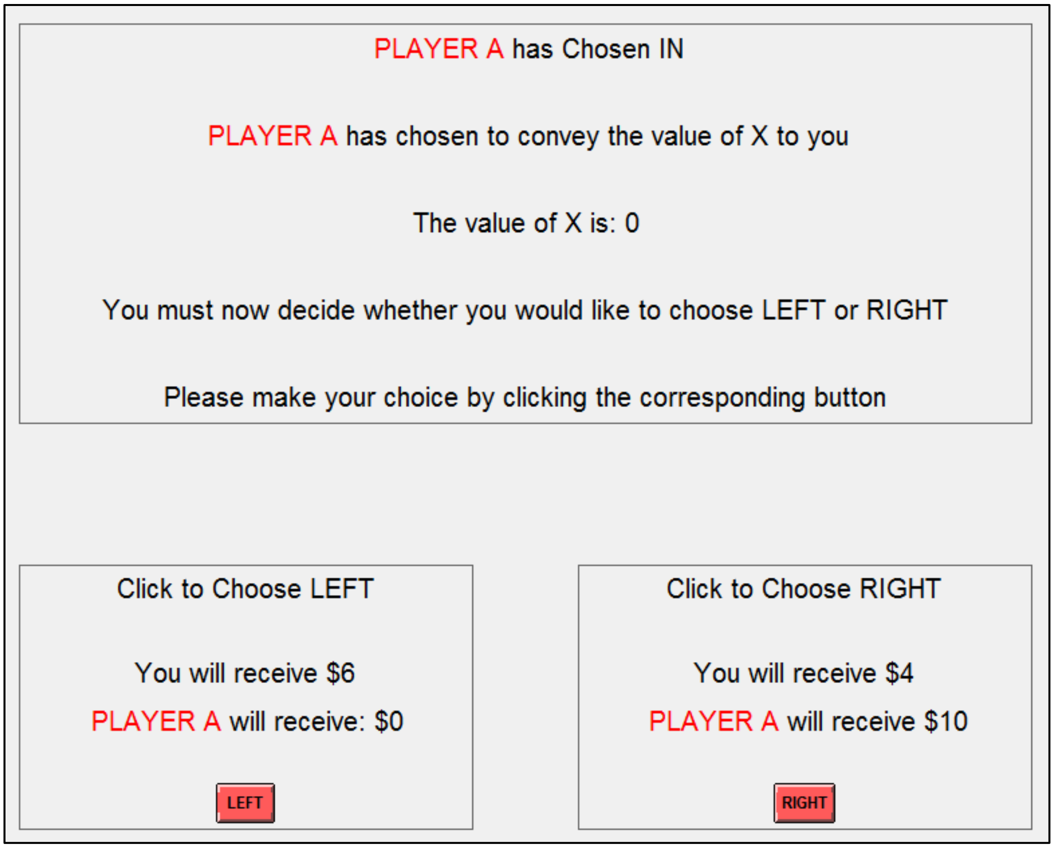
You must now choose whether you want to convey the value of X to PLAYER B before PLAYER B chooses between IN or OUT

Please make your choice by clicking the corresponding button

Click this button if you DO NOT want to convey X = 0 to PLAYER B

Click this button if you DO want to convey X = 0 to PLAYER B

Player B’s Conditional *Left/Right* Decision (X Known)



PLAYER A has Chosen IN

PLAYER A has chosen to convey the value of X to you

The value of X is: 0

You must now decide whether you would like to choose LEFT or RIGHT

Please make your choice by clicking the corresponding button

Click to Choose LEFT

You will receive $6

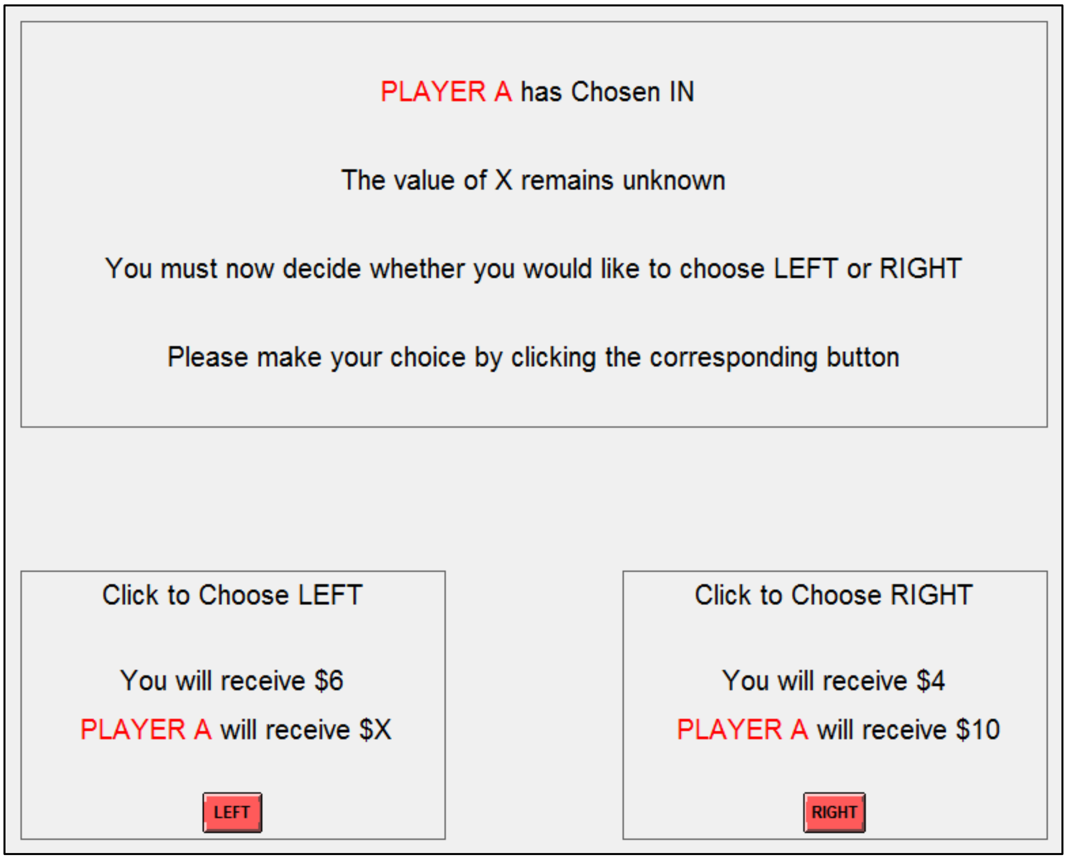
PLAYER A will receive: $0

Click to Choose RIGHT

You will receive $4

PLAYER A will receive $10

Player B’s Conditional *Left/Right* Decision (X Unknown)



PLAYER A has Chosen IN

The value of X remains unknown

You must now decide whether you would like to choose LEFT or RIGHT

Please make your choice by clicking the corresponding button

Click to Choose LEFT

You will receive $6

PLAYER A will receive $X

Click to Choose RIGHT

You will receive $4

PLAYER A will receive $10