Apologies as Signals: with Evidence from a Trust Game

Benjamin Ho[*]
Cornell University

Revised: July 2010

ABSTRACT

Apology is a social institution that restores frayed relationship not only in daily life but also in the domains of corporate governance, medical malpractice litigation, political reputation, organizational culture, etc. The theory shows that in a general class of moral hazard games with imperfect information about agents with two-dimensional type, apologies exhibit regular properties—e.g. apologies are more frequent in long relationships, early in relationships, and between better matched partners. A variant of the trust game demonstrates that communication matters in a manner consistent with economic theory; specifically, the words "I am sorry" appear to select equilibrium behavior consistent with the theory's main predictions.

Keywords: apologies, remorse, signaling, trust game, attribution theory

JEL Classification: C72, D82, L14, D23

[*] Author can be contacted by e-mail at bth26@cornell.edu by phone at 650-867-8270. Thanks to Edward Lazear, Douglas Bernheim, John Roberts, Chip Heath, Andrzej Skrzypacz, Emeric Henry, Florian Ederer, Luz-Marina Arias, Ted O'Donoghue, Justin Johnson for extensive comments, and to seminars participants at Stanford, Cornell and MIT Sloan. All remaining mistakes are mine. My apologies.

1 Introduction

> Paul needed a partner to expand his business. Several weeks ago, he made an appointment with his friend Amy to discuss details. Paul arrived on time, but Amy showed up one hour late. Paul is angry but after Amy apologizes profusely, Paul readily forgives and they have a productive meeting. They set a time for the following week, whereupon Amy is again late, again apologizes, and again is forgiven. On the third week, Amy is late once more, and at this point Paul is fed up. Talk is cheap, why do apologies have any meaning?[1]

Firms spend millions of dollars making amends with customers, electoral outcomes can shift from a single apology, peace treaties get derailed by the absence of apology, stock prices shift when a CEO apologizes, and millions of dollars in medical malpractice payments depend on whether doctors say the words, "I am sorry." While the popular press often asks why people do not apologize,[2] the mechanism of how apologies work is largely unstudied: When and why do people and firms apologize? What are the costs and benefits? Why does the benefit of deteriorate with use? If apologies mend relationships and talk is cheap, why is it sometimes hard to apologize? In an interconnected world where economic actors are embedded in a network of relationships, apologies act to restore frayed connections. This project provides an economic framework for understanding a social institution that is presumed to be based on emotion and validates the framework using results from the psychological literature and from the laboratory.

I focus on interactions where exogenous factors make payments infeasible, either due to legal reasons (e.g. a politician cannot bribe the electorate) or from norms (gifts[3] between friends tend to be limited to symbolic gestures). Also, I focus on principal-agent interactions, where payoffs depend on private information about the agent's two-dimensional type.

I do not deny that apologies may be driven by emotions like guilt, but I do not rely on psychological assumptions to allow for applicability to potentially highly rational actors such as firms, politicians and governments. If emotions like shame and guilt are the true motivation of apologies, this paper can address why these emotions arose, and why we pass on notions like remorse to our children.

Avoiding functional form assumptions, I propose a model general enough to encompass a broad class of principal-agent interactions—i.e. any game where an agent of unknown type performs a task for a

[1] The story is true, but the names and details have been changed to protect the "innocent."

[2] For example:" For fallen bankers, sorry may be the hardest, and smartest, word " International Herald Tribune, October 22, 2008; "Who should apologise to whom, for what and how?" Economist Oct 2, 2008, "Voters Hearing Countless Ways of Saying 'Sorry'" New York Times, Sept 1, 2006, "Learning Words They Rarely Teach in Medical School: 'I'm Sorry'" New York Times, July 26, 2005; "Australia Apologizing to Aborigines" Feb 13, 2008, "Video: I'm so, so sorry!" CNN Jan 27, 2006; "Being President Means Never Having to Say He's Sorry" New York Times, Oct 12, 2004; "Dear Economist…" Financial Times, June 8, 2007.

[3] There is a gift giving literature where agents choose inefficient gifts in order to signal type (Camerer, 1988; Prendergast and Stole, 2001, etc.). My approach here differs by being specific to the apology context in that it includes a moral hazard component. An apology occurs only after a transgression, whereas the traditional gift giving literature is concerned only with incomplete information.

principal with the possibility of repeated interaction—as well as a broad class of different types of apology—i.e. any communication that is designed to restore a broken relationship. In all such games, the model predicts several universal regularities that should be empirically observable: among them, apologies should be observed more frequently in longer relationships, when there is more uncertainty in the relationship, when the agent is a better match for the principal, and when the principal has more outside options. The model shows that apologies have competing effects on welfare (i.e. firm profits). Environments that allow apologies provide more information to principals and thus are welfare enhancing due to improved match quality. However, the ability to apologize worsens moral hazard. Similarly, the ability to apologize improves agent welfare via higher match quality, but this ability also burns money.

Important policy questions can also be addressed. For example, the emergence of "I'm sorry" laws in Texas, California, Massachusetts, Florida and other states make apologies inadmissible lawsuits involving medical malpractice. Apologies are also a prominent issue for politicians; the popular press constantly asks why politicians never apologize. By formalizing systems that have previously only been informally discussed, this paper clarifies these issues. I return to these applications along with connections to the psychology and sociology literature in the Discussion.

Though the model is designed to be as broad as possible, committing to an economic framework presupposes two attributes inherent to economic modeling. First, apologies are forward looking, not backward. The decision to apologize is primarily driven by the apology's impact on future payoffs rather than a guilt driven reaction to past events. Second, the role of the apology is external rather than internal. The primary motivation for giving an apology is to influence the beliefs and behaviors of others, rather than a response to internal conscience. These suppositions are testable and I find support for both in the lab experiment.

In the economic context, an apology is a costly signal that occurs after any game with moral hazard which restores relationships. Apologies have value not just because it is cheaper for good types to apologize or that good types get more value from the relationship (as in Spence, 1974) but also because good types fail in different situations than bad types. Apologies have signaling value even if single-crossing is not met. The value of an apology is found to be proportional to the cost; apologies without cost have no value at all, though I also show how the theory serves as a reduced form for many cheap-talk models. I test the theory with a novel trust game experiment. When the phrase "I am sorry" is used within a repeated trust game, the phrase serves as a signal in line with each of the theory's main predictions. Other phrases yield no significant effects.

An apology is defined as any act of communication "expressing regret or asking pardon for a fault or offense" (American Heritage Dictionary, 2000).[4] A mistake for *homo economicus* is the difference between what is ex post optimal and what is ex ante optimal due to a random move by nature. Though an apology is a signal sent by an agent to a principal, one should think of the players as part of a larger community where the principal needs a task accomplished and solicits the agent. Agents are ordered by how well suited they are to the principal's needs, i.e. their match quality. In a standard principal-agent framework, the match quality is the agent's cost of effort. In a political game, the quality would be alignment of ideal point. In a simple divide the dollar game, it would be a Fehr-Schmidt (1999) fairness or Becker (1976) style altruism parameter.[5] Social psychologists would call the agent's match quality, the agent's *disposition*, and I will speak of good dispositions and bad dispositions.

In the example above, the principal, Paul, partners with the agent, Amy, based on his beliefs regarding her punctuality, i.e. her match quality. During the first two apologies, Paul is willing to attribute the lateness to justifiable random events, e.g. traffic. After the third time, Paul concludes that he was wrong with his earlier attributions and ends the relationship.

In the following Section, I present the base model of costly apologies which provide the paper's main results. Section 3 explores cheap apologies and gives examples of models of cheap apologies that fall within the basic framework. Section 4 presents the experimental evidence. Section 5 relates the model to the broader psychological and sociological. Section 6 concludes.

2 Base Model

I begin analysis with two periods, but consider dynamics with more periods in section 2.2. Before the first period, nature sets the stable component of agent's type, $\theta \in \{\theta^G, \theta^B\}$ [6] with prior $p = \Pr[\theta = \theta^G]$. Recall that θ represents the agent's match quality or *disposition*. Then at the beginning of each period, nature sets the agent's changeable component of agent's type as $\omega \in \Omega$, with probability distribution $F(\omega)$ and Ω is a finite set.[7] We will refer to ω as the agent's *situation*. The agent's disposition is the

[4] There is an alternative use of the term "I'm sorry" in the English language where no fault is acknowledged, e.g. "I am sorry your grandmother died." By the above definition, this is not strictly speaking an apology, as the agent is not taking responsibility for the bad outcome. However, the ideas have somewhat become conflated, and thus I explore these "partial apologies" in the appendix.

[5] One might call this parameter, *sympathy*, see also Sally (2001, 2002). The American Heritage Dictionary (2000) defines sympathy as "A relationship or an affinity between people or things in which whatever affects one correspondingly affects the other." In Becker's model, a high altruism parameter means high utility for the other yields high utility for ones self.

[6] Results hold for a continuous distribution of types, but the notation is rather unwieldy, since first order stochasticn dominance does not provide a complete ordering.

[7] You could think of this model as representing any type-space that is decomposable into a dimension that is fixed across time, θ, and a dimension that is iid across time, ω.

component of type that is stable across periods. The agent's situation is the component of type that is uncorrelated across periods. The agent knows both dimensions of her type, the principal knows neither.

Then, through a production technology, $y(\theta, \omega)$, the agent produces output that yields payoff y for the principal as a function of her type, (θ, ω) where $E_\omega y(\theta, \omega)$ is increasing in θ. The model admits any production technology where agents can be ordered by a parameter, θ, which measures how well they produce for the principal.[8]

After production, both players observe the realized outcome, $y(\theta, \omega)$, at which point, the agent can choose whether to apologize or not, $a \in \{0,1\}$, at cost $c(a, \theta, \omega)$ (cheap talk variants are considered in Section 3). The principal updates his posterior beliefs about the agent's type, $b(a, y, p)$, using the realized outcome, y, and the apology, a.

$$b(a, y, p) = \Pr(\theta = \theta^G \mid a, y) \tag{1}$$

Finally, at the end of each period, the principal learns p_{out}, the probability that the outside option is a good type, drawn from a density function $G(p_{out})$. The principal chooses either to stay with the same agent or to take his outside option. If the principal stays with the same agent, the second period is played with a new draw of ω (θ remains the same). If the principal chooses the outside option, he gets a new agent with a new θ. The principal's utility is simply the sum of his payoffs from each period:

$$U_P(y) = \sum_t y_t(\theta_t, \omega_t) = y(\theta_1, \omega_1) + y(\theta_2, \omega_2) \tag{2}$$

Thus the principal maximizes his utility by interacting with higher typed agents, so he chooses the current agent over the outside option if the posterior probability that the current agent is a good type is higher than the probability the outside option is a good type. The principal's action as a function of his posterior, $\delta(b)$, is thus

$$\delta(b(a, y, p)) = \Pr[b(a, y, p) > p_{out}] = G(b(a, y, p)) \tag{3}$$

The utility of the agent is linear in three components: the agent's utility from production, the cost of apologizing, and the agent's discounted payoff from future interactions with the principal:

$$U_A(a \mid \theta, \omega) = u(\theta, \omega) - c(a, \theta, \omega) + v(b(a, y, p), \theta) \tag{4}$$

[8] This broad definition admits a broad class of situations/games where the apology game would be applicable. This reduced form specification of production abstracts away from potential moral hazard. Examples of production technologies that allow for agent moral hazard yet yield the same reduced form are given in Appendix B.

To make sure the agent wants to stay in the game, I normalize $u(\theta,\omega) > 0$. I normalize the cost function such that $c(0,\theta,\omega) = 0$ and let $c(1,\theta,\omega) > c(0,\theta,\omega)$. If the agent stays in the relationship, she gets payoff $u(\theta,\omega_2)$ in the second period, if she is "terminated" she gets payoff 0.[9] Utility in the two period model would be:

$$U_A(a \mid \theta,\omega) = u(\theta,\omega_1) - c(a,\theta,\omega_1) + \delta(b(a,y_1,p))u(\theta,\omega_2) \quad (5)$$

I consider pure strategy Perfect Bayesian equilibrium.

2.1 Analysis of Base Case

Consider first the game if apologies are disallowed. The agent produces. The principal updates his beliefs about the agent's type based on the output. The principal compares this belief to his outside option, and chooses to retain whichever agent he believes is better. Apologies give the agent the opportunity to provide an additional signal of her ability. After the outcome of production is observed, both principal and agent know what conclusions the principal will draw. The agent will apologize if and only if the apology shifts the principal's beliefs sufficiently to offset the cost of the apology.

In psychological terms, the principal either *attributes* a bad outcome of the production phase to the agent's disposition, i.e. a low θ, or to the agent's situation, i.e. a bad draw of ω. Psychologists study this type of dilemma using attribution theory (Ross, 1977, etc.). An apology is the agent's attempt to shift the principal's attribution of a bad outcome from the agent's disposition to the agent's situation. See the Discussion for details on the link to psychology.

Proposition 1: *In the apology game with c(a,θ,ω) increasing in a, v(b,θ) increasing and continuous a.e. in b, and Ey increasing in θ, any pure strategy Perfect Bayesian equilibrium with beliefs about θ as the state variable where both apologies and non-apologies are equilibrium outcomes at a given y has the following properties:*

a. *Principal beliefs b are weakly increasing in apologies, a, outcome, y, prior, p and agent's type θ.*

[9] People do apologize even in situations where we never expect to encounter the counterparty again. One could argue that we do this for the same reason we tip, that we behave, "as if" we will meet again, or that we follow some behavioral heuristic or script (Kahneman and Tversky, 1974; etc.). *v(b,θ)*, can also be thought of as psychic utility that stems from wanting to be liked.

b. *Probability that an agent apologizes,* $\Pr[a=1 \mid y]$, *is weakly increasing*[10] *in agent's type* θ. *The probability an agent apologizes approaches zero for extreme values of* p *and* y *and maximized for intermediate values.*

Proof in the Appendix. ■

Apologies always help strengthen/repair the relationship. Recall that $b(a, y, p)$ indicates the strength of the relationship. Furthermore, the stronger the relationship before the interaction, the stronger the relationship will be after the interaction. Better outcomes strengthen the relationship and in expectation, better types are liked more. As for apologies, good types necessarily apologize more often than bad types for a given situation. The likelihood of apologies is maximized when there is more uncertainty in the relationship, and for intermediate outcomes.

We can apply this proposition to the example between Paul and Amy. Amy produces timeliness, $y(\theta, \omega)$, each period. Each time Amy shows up late, Paul shifts his beliefs that she is a good match, $b(a, y)$, downward. However, each time, Amy's apology shifts his belief upward. So long as Paul's confidence in Amy is higher than his outside option, $b(a, y) > p_{out}$, Paul retains Amy as a partner.

The results from this section can be summarized by the following graphs:

(Insert Figure 1 about here)

The first graph plots the probability of tendering an apology given some outcome, $y(\theta, \omega)$, against the principal's prior, p. Good agents always apologize more, and apologies are most likely for both types at intermediate values of p. In other words, we should see more apologies when the two parties are better matched, and when there is more uncertainty in the relationship.

The second graph plots the principal's posterior, $b(a, y)$ against the principal's prior. The success line is for a high value of y without an apology. The "agent apologizes" line is for a bad outcome of y where $a = 1$. The "agent doesn't apologize" line is for a bad outcome of y where $a = 0$. Apologies are most effective when the principal is most uncertain about the agent.

Agents do not apologize for high outcomes. What determines the cutoff for when an apology is tendered is given by the slope of the continuation payoff which depends on the distribution of outside option and the value of future interactions. Agents are more likely to apologize when principals have better outside options and when there is greater scope for future interaction.

Note that the Proposition does not place any special restrictions on the cost or the payoff function. However, it does depend on existence. Conditions for existence are given as follows:

[10] The relationship would be strictly increasing, except for corner solutions where the probability of apology is either

Proposition 2: *A Perfect Bayesian equilibrium in which apologies are used with positive probability given some outcome* y *exists if and only if the set of states that yield such a* y *where good types find it beneficial to apologize is weakly more probable than the set of states that yield such a* y *where bad types find it beneficial.*

Proof in Appendix. ■

Restated symbolically, this proposition states that a separating equilibrium exists when the following condition holds:

$$\begin{aligned}\Pr\left[c(1,\theta^G,\omega) \le v(b(1,y,p),\theta^G) - v(b(0,y,p),\theta^G) \mid \theta^G, y\right] \ge \\ \Pr\left[c(1,\theta^B,\omega) \le v(b(1,y,p),\theta^B) - v(b(0,y,p),\theta^B) \mid \theta^B, y\right]\end{aligned} \tag{6}$$

This condition is not the most intuitive, so I consider the condition with θ and ω separately. Apologies can work for any one of three reasons. To see the first two, we eliminate the cost function's dependence on ω in Equation (6), the uncertainty goes away and all that remains is the familiar single crossing property.

$$\begin{aligned}v(1,\theta^G) - c(\theta^G) - v(0,\theta^G) - 0 \ge 0 \ge \\ v(1,\theta^B) - c(\theta^B) - v(0,\theta^B) - 0\end{aligned} \tag{7}$$

Apologies work if the cost of apologizing is lower for higher types. Or, apologies work because higher types expect higher future benefits from staying in the relationship.

Alternatively, since costs can vary by the situation, ω, the model allows apologies to work if different types fail in different situations. Unlike one-dimensional models, dependence on θ can be removed and signaling can still be effective. Removing θ the condition becomes:

$$\begin{aligned}\Pr\left[c(\omega) \le v(b(1,y,p)) - v(b(0,y,p)) \mid \theta^G, y\right] \ge \\ \Pr\left[c(\omega) \le v(b(1,y,p)) - v(b(0,y,p)) \mid \theta^B, y\right]\end{aligned} \tag{8}$$

An apology equilibrium still exists so long as the set of situations, Ω_G^y, that a good type fails in contains more low cost ω's than the set of situations, Ω_B^y, that bad types fail in. Given the right technology, apologies work even if single crossing is not satisfied. For example, if we assume it is more costly to maintain a lie than to tell the truth, then it is easier for Amy to excuse her tardiness by citing bad traffic, if there really had been bad traffic.

---

zero or one.

Note, this is the unique separating Perfect Bayesian equilibrium, since there is a unique best response strategy in each sub-game. As is typical, a pooling equilibrium also exists, which will be considered when we consider welfare and in the experiment.

2.2 Examples of Production Technologies

Since apologies occur in a wide variety of social situations, the model is designed to accommodate a broad class of moral hazard games where agents may take actions that require an apology. Appendix B demonstrates how any moral hazard game where an agent might have a reason to apologize is equivalent to some $y(\theta, \omega)$ specification so long as a supermodularity condition is satisfied. Appendix B provides detailed examples of several moral hazard games (principal-agent games, political games and dictator games) and how various reasons for apologizing (changing moods, a change of heart, or unforeseen circumstance) can be modeled.

2.3 Examples of Cost Functions

The analysis of this paper is designed to hold for fairly general production technologies. So long as moral hazard is properly maintained and the existence conditions are satisfied, a wide array of cost mechanisms can be represented.[11]

Relationships can be restored using a wide variety of different apology mechanisms. When people use the words "I am sorry," they mean different things in different circumstances. An apology can represent a tangible cost, such as in emotional pain,[12] time spent, flowers purchased, or an acknowledgment of responsibility with third party legal ramifications. An apology can represent a commitment to work harder in the future. Or an apology can represent a loss of status as is often the case with politicians. An apology could also entail the giving of an excuse where a false apology must face sincerity detection.

Some argue that an excuse is not the same as an apology. In fact, the word apology derives from the greek word *apologos* or story, and came into usage to describe an account used to excuse a transgression (Tavuchis, 1991). In more common contemporary usage, the Merriam-Webster (2008) dictionary defines apology firstly as "1a: a formal justification, b: excuse" and only secondly as "2 : an admission of error … accompanied by an expression of regret."

[11] The cost paid for an apology is given exogenously instead of being chosen by the agent. Allowing a to be a continuous variable would not substantially change results.

[12] The costs can be entirely behavioral, such as through norms of shame that surround the very act of giving an apology, mitigated by alleviation of guilt. Tavuchis (1991) argues that society makes it painful to apologize through norms of social disapprobation, because the pain is what gives apologies meaning. The model here is consistent with this view of apologies. Again, although there is considerable psychological and sociological evidence that such mechanisms of shame and guilt are at work, I focus on economic explanations.

Section 3 shows that so long as the assumptions hold, all of these apology mechanisms—whether the apology operates through excuses, commitments, status, empathy, or emotional pain—still share the properties derived in this section. A more structural view of each mechanism does yield more precise insight, a full discussion of which is left to Appendix C.

2.4 Costs and Welfare

There are three main exogenous parameters of the model, the agent's type, the principal's prior beliefs, and the cost function. The impact of each on the principal's welfare is straightforward.

Proposition 3: *In the apology game with $c(a,\theta,\omega)$ increasing in a, $v(b,\theta)$ increasing and continuous a.e. in b, and Ey increasing in θ, any pure strategy Perfect Bayesian equilibrium using beliefs about θ as the state variable where both apologies and non-apologies are equilibrium outcomes at a given y has the following properties:*

a) *The change in the principal's beliefs—i.e. the effectiveness of the apology—is weakly increasing in the cost of the apology.*

b) *The principal's welfare is weakly increasing in the agent's type, θ.*

c) *The principal's welfare is weakly increasing for higher prior, p, when the agent is good; The principal's welfare is weakly decreasing in prior p, when the agent is bad.*

d) *The principal's welfare is maximized for intermediate costs of apology.*

Proof in Appendix. ■

High cost apologies provide more information for the principal, but only if they are used. However, for costs that are too high, the condition in Proposition 2 is violated, and the principal gets no information at all. Thus welfare is maximized for intermediate costs. More generally, the institution of apologies provides more information for the principal and thus increases welfare. Moreover, when production depends on effort, there is an ambiguous effect on welfare, as the institution of apologies also increases the ability to shirk.

The institution of apologies tends to decrease agent's welfare. Allowing the agents to apologize gives them an opportunity to burn money to signal, an opportunity that agents are compelled to take.

2.5 Dynamics of the base case

The number of periods in the apology game can be repeated for T periods. The posterior in each period becomes the prior in the next.

Proposition 4: *In any given period t of a T period apology game, the results of Proposition 1 hold for each stage game if the continuation value in each period is increasing in the priors.*

Proof in Appendix. ■

The assumptions that the continuation value is increasing in the prior will hold so long as the prior is sufficiently high, which is a plausible assumption so long as we assume that the principal initially matched with the agent associated with the highest prior.[13] Also note that in the long run, since $b > p$ after interactions with good types and $b < p$ after interactions with bad types, the principal's beliefs become more accurate with time.

To see how these results can be utilized let us return to the motivating example. Paul is choosing between two potential partners Amy and Alice, and wants a partner of a good type. Paul's prior is that Amy is more likely to be a good type than Alice, so he partners with her. After the first instance of Amy's tardiness, Paul has a dim view of Amy's type, but her apology restores his beliefs closer to but still below his initial prior. High enough, though, so that Amy is still more likely to be better than Paul's outside option, Alice. However, a second failure and apology lowers Paul's confidence in Amy further. The third failure lowers Paul's belief in Amy's type so that even after an apology, Paul's belief that Amy is good is now lower than Paul's prior that Alice is good. Paul ends the relationship with Amy.

Successes strengthen relationships. After failures, apologies restore relationships but only imperfectly. An apology after one failure may restore the relationship sufficiently to continue the game, but a succession of failures can leave b so low that an apology cannot save it.

3 Cheap Apologies

The model presented in the previous section is effective at explaining many of the characteristics observed in apology interactions. However, its reliance on an exogenous cost function leaves the modeling unsatisfying. In this section, I return to the question originally posed: if talk is cheap, why do apologies have any meaning? I now assume that an apology is simply a message $a \in \{0,1\}$, and its meaning determined in equilibrium. The models described in this section conform to the assumptions from Section 2 and thus shares the properties expressed in Propositions 1 through 4, but each unpacks the black boxes that the cost, $c(-)$, represents. Note that some additional mechanism is necessary:

Proposition 5: *If the cost of apologies in the base model is set to zero, there is no Perfect Bayesian equilibrium in which apologies reveal information.*

[13] We have established that in the final period, higher posterior yields higher payoffs. In each stage game, proposition 1 tells us that the frequency of apologies is decreasing in the prior for priors that are sufficiently high, so we just need the prior to be high enough so that total cost of apologizing is decreasing in the prior.

This proposition follows because if apologies were used differentially by the two types, one message would shift beliefs higher. Then both types would send the same message.

I propose instead several modifications that endogenize the cost in the base model. Though each variation can be modeled in reduced form as a cost function, it is instructive to consider how introducing different game mechanics can give meaning to cheap apologies.

Third party enforcement: The simplest mechanism creates a tangible cost by introducing a third party. Assume a legal system that can impose a fine or legal sanction on the agent. The probability that the sanction is imposed depends on the agent's guilt which depends on the agent's actions and his situation. The cost of the apology comes from the fact that the apology can later be used as evidence in court against the agent.

Lying: The word apology derives from the Greek for story. An apology is an example of account giving or excuse giving. The cost of account giving is based on the probability the lie will be found out. Amy claims she was late due to traffic, but there is some chance Paul will hear a traffic report and catch her lie. Or alternatively, many argue that humans are evolved to be poor liars and dishonesty can be detected by facial cues (Ekman, 1969; Frank, 1989). In either case, the possibility exists that a lie will be found out, and the cost represents the punishment associated with lying, determined socially outside of this model.

Formally, one could endogenize the cost of lying or excuse giving by having an apology be a message that reports the state of the world ω that is initially unobserved by the principal. However, with some probability, false reports are caught which leads the relationship to be severed. Honest reports are never punished and thus costless. A good agent would report ω truthfully as it is costless to do so. If a bad agent reported ω truthfully, the principal knowing $y(\theta, \omega)$ and ω could back out the agent's type, thus if the bad agent wants to be perceived as a good agent, she must lie. The cost of lying depends on the probability of getting caught which depends on ω and also the incentive to lie which depends on θ. A good agent would always be able to tell the truth and incur zero cost for such an apology.

Commitment: This possibility could be called the "I'm sorry I won't do it again" apology. It depends on the principal's ability to commit to firing an agent known to be good. The resulting "fool me once, shame on you, fool me twice, shame on me" equilibrium demonstrated in the appendix uses the apology as a contract, a promise to never do it again. The principal maintains incentive compatibility by punishing a failure that follows an apology more severely than if there was no apology at all.

Status: A common reason given that it is difficult to apologize is that an apology entails a loss of status, a shift of the principal's beliefs in the status dimension. One could argue that the reason women apologize more than men arises from evolutionary pressure that made status more important for men as men need status to compete for mates. Since status matters relatively less for women, they can apologize more. The

same can be said for Asian cultures relative to Western ones. If Asian cultures value group preference alignment more than individual ability, then apologies will be more prevalent in Asian cultures.

Status concerns are especially critical for political apologies. Tiedens (2001) experimentally demonstrates that even though politicians gain approval and liking by apologizing, an apology causes the politician to lose status as measured by respect or willingness to re-elect.

Status loss could be modeled with a renegotiation proof contract based using two-dimensional type. Up until now, the point of an apology was to shift the principal's attribution of the cause of a bad outcome from an internal and controllable quality of the actor, to an external and uncontrollable quality of the environment. Lee and Tiedens (2001a) find that even a successful shift of attribution may not be good for the agent, if the agent is expected to have control over the environment. I construct an equilibrium where a principal can offer a menu of tasks that depend on both intentions and ability. Agents who apologize receive new tasks correlated to intention, while agents who do not apologize receive new tasks correlated to ability.

The details of these mechanisms along with a model of empathy-based partial apologies based on information partitions are relegated to the Appendix B.

4 Experimental Evidence

The theory presented above constructs a formal model of apologies. The intent is to understand apologies in broader contexts such as political campaigns or medical malpractice litigation. Thus, it is useful to validate the model's predictions with experimental evidence in order to gain more confidence in the theory's validity. Experiments also allow the exploration of additional regularities of behavior, specifically other features of the dynamics of apologies, implications for social welfare, and the effect of participant demographic characteristics.

4.1 Background: The Trust Game

The trust game provides a natural test bed for the theory. It is a two player game where a task must be performed by one player—the agent—for the benefit of the other—the principal. It is a game where responses differ across individuals, and the actions of some individuals have the natural interpretation as being nicer than the actions of others, a situation that lends itself to apologies. Furthermore, the structure of the trust game provides clean measures of the five main variables in the model: type, prior, outcome, apology and posterior.

The trust game is also useful because the typical behavior observed does not conform to traditional restrictive models of rational behavior. The application of the model to the trust game demonstrates an important feature of the apology model. It does not depend on the functional form of any psychological assumptions used. Instead, the theory applies to a broad class of them.

The trust game as first proposed by Berg, Dickhaut and McCabe (1995) is a one shot game with two players, a Principal/Investor (Paul) and an Agent/Trustee (Amy). Paul is given some number of tokens, typically ten. Paul can choose to keep any number of tokens; the rest get invested with Amy, who receives three times the number that Paul invested. Amy then chooses to keep any number of the tokens, the rest are returned to Paul.

In the Berg et al. (1995) study, the modal response was for the Principal to send half, or five, tokens to the Agent, giving her a resource pool of 15 tokens to work with. The modal response of the Agent was to return five back to the Principal, providing an average net return on investment for the principal of 0%. Following Berg et al. (1995) study a large literature shows among other things that behavior depends on culture (Koford, 1995; Ensminger, 2000), and by gender (Chaudhuri and Gangadharan, 2007), thus the "type" of the person one interacts with matters. Also, there is some debate as to whether behavior in the trust game is driven by reciprocity or altruism (Gneezy et al. 2000; Cox, 2002; Chaudhrui and Gangadharan, 2003). The model presented holds for either. Charness and Dufwenberg (2006) find that communication of promises increase trust and argue that guilt affects trust behavior.

More directly, Schweitzer, Hershey and Bradlow (2006) analyze the impact of apologies in a repeated trust game scenario. Their treatment uses deception and uses an actor in the role of the agent. The advantage here is that I have observations for both the actions of the principal and the agent. Furthermore, the game is constructed so that theoretical predictions can be made in advance, and therefore the model can be directly tested.[14] They find that though an apology restores trust, a failure after an apology is punished especially severely. Ohtsubo and Watanabe (2009) present several experiments using hypothetical vignettes and a dictator game to demonstrate that the perceived sincerity of an apologies is increasing in the cost, while Abeler et al. (2009) find in a field experiment that a cheap apology is more effective than a payment.

4.2 Experimental Design

The experiments in this paper use a repeated and noisy variant of the trust game programmed using z-Tree (Fischbacher, 2007). In each session, there are 8-12 subjects, with exactly half assigned the role of principal and half assigned the role of agent (See Appendix D for the instructions). Each principal is randomly paired with an agent with whom he plays a noisy trust game for ten periods. After each period, players learn the realized payoff for both players, and the game repeats. After ten periods, that particular treatment ends. Each principal is re-matched with a different agent from the same session until all possible pairings are made.

[14] Notably, the Schweitzer et al. (2006) analysis provides empirical confirmation for the "fool me once…" prediction from the dynamic cheap talk model in the appendix.

Given prior literature, I did not expect, nor did I find that the finite and certain number of periods leads to unraveling. Indeed, the lack of unraveling in a setting where theory would traditionally expect it highlights an important feature of the apology theory: utility functions based only on more than monetary gain are allowed in the model.

The noisy variant of the trust game is similar in structure to the standard form. The principal is given 10 tokens of which he can choose any integer number greater than zero to keep for himself.[15] The rest are entrusted to the agent. As in the standard game, the agent receives the number entrusted by the principal, multiplied by three which goes to form the agent's resource pool. The agent can also choose to keep any number of tokens for herself, with the remainder returned to the principal. However, instead of being directly returned, the noisy variant introduces uncertainty. For each token that the agent returns, the principal gets a cumulative five percent chance of "project success." A successful project yields the principal 20 tokens. Thus if 15 tokens are returned, the principal gets 20 tokens with probability 0.75. Returning more than 20 does not help. Since the return of more than 15 tokens was never observed in past experiments, the game is essentially the same, in expectation, as the canonical trust game, except that the principal is unaware of the number of tokens the agent actually returned.

At the end of each period, the agent is told the outcome of the project, and given the opportunity to send a message. The only message she is allowed is a message that reads "I am sorry."[16] The players were told the communication cost associated with sending a message before each treatment and were randomized across treatments. Communication costs of 0, 5, 10 and 15 were equally likely.

Subjects were primarily undergraduates recruited through the Stanford Graduate School of Business Behavioral Lab. Advertising for the experiment was restricted to those who had never participated in an experiment before. Of the 58 subjects recruited for the base study (out of 110 recruited across all conditions), 53% had never been in an experiment involving deception. 57% were female. 44% were white, 20% were Asian, 26% were other. 55% had never taken a psychology class and 50% had never taken economics. 22% thought that deception might have been involved. Regressions using these data showed no significant effect on observed behavior except non-white, non-Asians agents returned 17% fewer tokens, agents who had taken economics returned 11% more, and principals who had been in experiments before entrusted on average 1.5 tokens more.

Subjects were given an instruction sheet (see Appendix D) that detailed all aspects of the experimental design and encouraged to ask questions. They were assured that the norms of economics precluded deception from being used. A trial run was conducted in which the experimenter was available to answer questions. Principals and agents were seated in separate rooms to promote anonymity.

[15] Principals were not allowed to send zero so that agent behavior would be observed in every period.
[16] At the time, the software at the time made more open-ended messages unworkable.

4.3 Correspondence to Theory

This noisy variant of the trust game was chosen because it conforms well to the model, and the measurable actions provide good proxies for the variables in the model. Recall the five key variables of interest. Wherever averages are specified, I consider averages for a given player over the ten period treatment, as well as averages over the entire session.

- **θ**: The theory defines θ as the measure of type where higher types yield higher payoffs for the principal. In the trust game, two natural measures have this property: the average number returned by the agent and the average percent returned.
- **p**: The theory defines *p* to be the prior belief the principal has that the agent is a good type. How much the principal chooses to entrust to the agent is monotonically increasing in *p*. A principal who is just maximizing tokens would tautologically like to give more tokens to a higher type, because that is how higher type is defined. Thus conversely, the more tokens a principal entrusts, the higher his beliefs.[17]
- **y**: This is the output for the principal. It is 20 in case of project success, and zero in case of project failure.
- **a:** In the theory, *a=1* if the agent apologizes to the principal after a failure and *a=0* otherwise. This is simple to measure given my setup.
- **b**: This is the principal's posterior belief. Since the principal's beliefs are measured by how much he entrusts, the beliefs at the end of a period can be measured by how much he entrusts at the beginning of the next.

Note that in the interest of robust experimental design, it would have been preferable to assign type exogenously. However, an apology depends on prevailing social norms and it was supposed that an apology would be ineffective without a sense of moral obligation on the part of the agent, and that moral obligation would not exist if a low payoff for the principal was due to something assigned by the experimenter. I will present evidence that behavior does depend on some external sense of morality by showing that it is not merely the structure of the game that produces the signaling equilibrium, but instead, the words themselves matter.

It should also be noted that it cannot be the case that the agent is merely maximizing tokens, otherwise standard backward induction unraveling results would be expected: no tokens would ever be entrusted, so no tokens would ever be returned. Since unraveling rarely occurs in the trust game, there must be some element of the utility function that causes at least some types (even just an ε fraction) to

[17] Mathematically, a simple functional form would have θ be the percent an agent is expected to return. Then a per-period profit would be $\pi = (10 - x) + E3x\theta$, so $\partial\pi / \partial x = \Sigma[3\theta - 1]b(\theta)$, so the principal should entrust

return a positive amount. One such set of types would include 1-ε token maximizers and ε "crazy types" who always return tokens as in Kreps, Milgrom, Roberts, and Wilson (1982). Alternatively, type could reflect differences in discount rates, or degrees of altruism in a model with social preferences. Regardless of what type reflects, the theory presented here encompasses all examples where types can be ordered by how much they return on average. This flexibility highlights the advantage of this theory as opposed to one that presumes a functional form.

Finally, I consider the existence conditions in Proposition 2. To keep things simple for the subjects, constant costs of communication is assumed,[18] so what is required is the standard single crossing condition. Whether single crossing actually holds depends on the functional form of utility. However, from Proposition 2, if apologies are observed, then the condition must hold presuming the other assumptions are also satisfied.

4.4 Results

The basic results reflect six sessions totaling 58 subjects, or 125 pairings of ten periods each. All of the main predictions from Proposition 1 were found significant at the 95% level.

4.4.1 Basic Results

(Insert Table 1 here)

Consider first Table 1 which lists summary statistics of the relevant variables broken down by communication costs. Apologies were used frequently, 38% when costs were zero, 15% when the cost was 5 tokens, and 3% when the cost was 15 tokens. On average, of the 10 tokens that the principals received each period, they entrusted about half in line with past trust experiments.

The remainder of this section is motivated by Proposition 1 of the theory, which makes nine predictions about the anticipated relationship between the variables.

First, the principal's posterior belief, *b*, should be increasing in each of the four other variables. The principal should entrust more in the following period if in the previous period, the agent apologized, *a*, the project succeeded, *y*, the agent is a better type, θ, or the principal had high beliefs and entrusted more to begin with, *p*. Table 2 shows that these predictions are all significantly borne out.[19] The addition of agent fixed effects strengthens the results.

---

everything for beliefs sufficiently high, and nothing for beliefs sufficiently low. Though adding risk aversion would give intermediate results.

[18] To be precise, I am assuming constant observable costs of communication. I interpret noise in the estimation procedure to come from some internal psychic/social cost of apology that I cannot observe.

[19] The tables are aggregated over the first 9 rounds of the game. The results still hold when considering any one round, though for rounds 9 and 10, they become insignificant due to end of game effects. F-tests cannot reject the null hypothesis that period dummies for the first eight periods in the regressions in Tables 2 and 3 are all the same.

(Insert Table 2 here)

The regressions reported here include all observations from the first nine periods and include period fixed effects. One might be concerned about autocorrelation though assuming an AR1 process does not materially change results. Similarly, running a separate regression for behavior in any given period (e.g. the ninth period) yields consistent and similar results though the significance is reduced for some coefficients given the smaller sample sizes.

The model's predictions for the likelihood of apology are more involved. Higher types are expected to apologize more often conditional on outcome and is seen in Table 3. Recall that the proxy for higher types are those agents who over the entire session return a higher percent of their resource pool. The theory also predicts that for extreme failures, the agent is also unlikely to apologize, but since project outcome is a binary variable, this was not observed. The theory predicts that apology frequency is maximized for intermediate priors. For intermediate priors, apologies have the most value since there is most uncertainty. The interior maxima predicted by the theory for the effect of priors on apologies are borne out by the data. The regression including fixed effects predicts the likelihood of apologies to be maximized when the principal entrusts 6 tokens.

As for predictions about project outcome, *y*, higher types yield higher project outcomes, which is true by definition.

(Insert Table 3 here)

Note that both the tendering of an apology and the outcome of the project are binary variables. The regressions were replicated using logit and probit analyses yielding similar results. Also, redoing all presented regressions clustering standard errors by principal or agent does not substantively affect significance.

4.4.2 Welfare and Dynamics

From Proposition 3, one would expect principal profits to be maximized for intermediate costs. Profits were significantly higher at intermediate communication costs in the pilot and in the majority of sessions, but two sessions with unusually high profits at costs of zero and low profits at costs of 5 negated that finding for the whole sample.

There is some evidence for the dynamic predictions of the model (see Appendix: Section 8.1). There is weak confirmation for the finding of Schweitzer et al. (2006), that apologies followed by failure should be punished more harshly than non-apologies followed by failure. Looking at the level of trust in trials where communication costs are zero, I find some evidence for the converse of the "fool me once, shame on you…" cheap talk equilibrium from Appendix C (see Equation (30)). Successes following non-

apologies appear to be rewarded more highly, to compensate agents who took an initial loss by not apologizing. However, all results of zero cost apologies were statistically insignificant.

4.5 Robustness Checks

To check whether apologies are forward looking and economically motivated by future payoff, or backward looking and psychologically motivated by guilt, I look at what periods apologies are used in. Apologies show a significantly negative time trend; the frequency of apology declines by 1.2% (t-value equal to 3.01) each period, with average frequency 21.5%.

(Insert Figure 2 here)

Also, if apologies are forward looking, then the agent's continuation value—defined as the agent's profits for the remainder of the game at the end of each period—should justify the cost of apology. Continuation values are consistently higher for agents who apologize than for those who did not. Regressing continuation value on the number of tokens an agent spends on an apology (including period fixed effects) each token spent yields the agent 1.37 tokens future profit; this result is robust to adding controls for prior beliefs and agent fixed effects.

(Insert Table 4 here)

Another backward looking hypothesis is that the agents are burning money in order to restore fairness. However, a fairness motivation implies that higher agent profits should increase the likelihood of an apology. Regressions not reported here show fairly precise insignificant results.

Xiao and Houser (2005) run an ultimatum game experiment where cheap messages with affective content do change the behaviors of the players. To explore this possibility, I conducted the noisy trust game again. Instead of restricting the message to "I am sorry," I instead allowed a choice of five messages: "I am happy," "I am sad," "I am neutral," "I am sorry," and "I am angry." Each of the messages except for "I am neutral" which was essentially never used, were used with about equal frequency, though overall, messages were used much less frequently, about a third as often as in the original game. Uncontrolled regressions show that other messages have somewhat similar effects on beliefs as "I am sorry" which continues to positively impact trust. However, once agent type is controlled for using average number returned, one finds that only "I am sorry" is significantly effective.[20] I would argue that though the theory stands the same, regardless of message, the specific "I am sorry" message triggers context cues in the participants to play the separating equilibrium while for other messages the

[20] Another trial where messages such as "I am happy" or "I am sad" are assigned to players yielded similarly insignificant results.

default response is to pool. Alternatively, equilibrium play in the apology game is governed by certain heuristically adapted strategies that are not triggered by other messages. A third possibility is that there are unobserved costs associated with the words "I am sorry" representing other proposed mechanisms such as commitment[21] or status loss, as modeled in the Appendix C. Clearly, there is something outside the context of the model that affects behavior. Though rational signaling can explain much of the variation, behavioral factors still play a role.

(Insert Table 5 here)

5 Discussion

5.1 Literature Review: Psychological and Sociological approaches

The psychology literature provides many properties of apologies, but largely ignores the mechanism. A typical reason given for why people apologize is limited to "negative affect alleviation." People feel guilty, and an apology removes guilt. This view dates to a Freudian model of behavior where humans have stocks of emotion—e.g. guilt—which when accumulated, causes distress until the person apologizes (Cialdini et al., 1975; Cialdini and Ascani, 1976).

Tavuchis (1991) has a sociological treatment of apologies in which he sees apologies as a social system designed to maintain relationships and establish membership in community. Tavuchis calls apology a social exchange, a device that paradoxically restores social order without amending the transgression. Nothing material has been exchanged, yet the relationship has changed. The pain of an apology is created by a social system of shame that accompanies it.

These psychological and sociological models of behavior while not consistent with economic rationality, may be accurate. However, one interpretation of the theory is that it is a mechanism that creates the cost function for an apology that maximizes social welfare. Society benefits if people apologize, but the act must induce shame to make the apology meaningful. Guilt helps encourage good agents to apologize. As in Frank (1989), culture provides devices such as guilt, remorse and shame, in order to facilitate optimal strategic play. The rational actor assumption can be used for modeling purposes using Friedman's (1953) "as-if" justification.

Attribution theory from social psychology offers a more useful albeit incomplete theory of apologies. Attribution theory concerns itself with settings where an outcome is observed, e.g. Amy is late for a meeting, that could have two possible causes: 1) dispositional, Amy is lazy and inconsiderate, or 2) situational, Amy was held up by unexpected traffic. Psychological theories about the fundamental attribution error predict that individuals attribute too much blame to the disposition of the actor and not

[21] The data is consistent with the "fool me once…" model of commitment in the appendix, but not significantly so.

enough to the situation (Heider, 1958; Ross, 1977; Jones and Nisbett 1972).[22] Weiner et al. (1987) applies attribution theory to the process of apologies, and in experiments, finds that apologies that attribute bad outcomes to external uncontrollable situations, ω, are more likely to maintain a relationship relative to apologies that attribute bad outcomes to internal controllable dispositions, θ. In the model presented here, an effective apology shifts the principal's attribution of the cause of the bad outcome from the agent's disposition to the external situation. The model demonstrates that in the absence of an apology, attributing a bad outcome to the principal's disposition is not an error, but a rational response to the available information.

5.2 Applications

Interpersonal Relationships: Apologies are a common occurrence in everyday life, particularly in the maintenance of friendships. Empathy is particularly important in this context. The model also addresses cultural differences in apologies—Asians apologize more than Americans (Takaku et al., 2001)—or gender differences—women apologize more than men (Gallup, 1989 in Tavuchis, 1991). What psychology literature exists focuses on experimentally validating stylized facts. An apology by the agent reduces the anger the principal feels toward the agent as well as the principal's desire to punish (Ohbuchi et al., 1989). In tasks where the agent is less responsible or where the offense is less severe, the apology is rejected less often (Bennett and Earwaker, 2001). Apologies are almost always accepted (Mullet et al., 1998; Bennett and Dewberry, 1994). Forgiveness occurs more often in closer relationships (McCullough et al, 1998). The results of all the psychological experiments I found in the literature were consistent with the findings in the model presented.

Organizations: The prevalence of apologies in various organizational settings is indicative of differences in task assignment, risk taking, turnover, conflict resolution, complementarities in production, etc. A model of apologies offers insight into cultural differences in organizational design. Lee and Tiedens (2001a) find that within an organization, when individuals in control make excuses for their behavior, they lose status. Proposition 5 explains that this status loss is necessary for apologies to be effective.

Corporate Governance: CEOs are responsible to their shareholders. When performance is low or scandal arises, should an apology be expected? Does an apology carry any weight? Lee and Tiedens (2004) find that certain kinds of attributions for past performance found in company annual reports—effectively apologies—can predict a firm's stock prices one year out. We demonstrate that the effectiveness of a corporate apology depends on the costs of the actions the firm undertakes, how the

[22] Though not addressed in this paper, principals in this paper rationally commit the fundamental attribution error, in order to give agents incentive to provide costly information.

costs of those actions compare to the actions of less favorable types, and the value of a continued relationship with the customer.

Politics: There is a stylized fact that politicians never apologize. Consider, Bush on Iraq, or Clinton on Lewinsky. Tiedens (2001) conduct an experiment where she constructs two videos by splicing interview footage, one where Clinton appears to apologize regarding the Lewinsky affair, and one where he appears angry. Subjects who saw apologetic Clinton liked him more, while subjects who saw the angry Clinton liked him less and complained about how Clinton never apologized. However, on questions of leadership, ability, and importantly, whether you would re-elect, the angry Clinton fared better. The result is robust to choice of politician and the crime. Appendix C explores these status based apologies in greater depth.

Governments: Governmental apologies occur either between the government and its people (e.g. South African apartheid, Japanese internment, or slavery in the United States), or between governments in international relations (the difference between the German and Japanese response to World War II). The South African Truth and Reconciliation Commission increased welfare by effectively lowering the cost of apology so that apologies can be made and relationships restored. Simple punishments would have left social relationships broken.

Litigation: One of the few areas of scholarly research that examines intensively the question of apologies is in the area of law. Apologies have an important impact on the outcome of cases. Unsolicited apologies can have an impact on conviction rates, as well as sentence and judgment sizes (Rehn and Beatty, 1996). Furthermore, even court ordered apologies appear to mitigate punishment (Latif, 2003). Ho and Liu (2009) employ the apology model in a differences-in-differences analysis of the impact of apologies on medical malpractice litigation rates.

5.3 Conclusions

Apologies are a social institution that can be understood using the standard economic toolkit. This paper provides a model that encompasses many of the myriad settings in which apologies occur. In each setting, an apology acts as a signal of the apologizer's fitness for future interaction. Conditions are established for the unique existence of such equilibrium. Credence can be given to these theoretical predictions because findings from a trust game experiment match the theory's predictions better than alternatives.

Taken together, the results here help explain behavior in many economic contexts, including product safety recalls, medical malpractice lawsuits, political campaigns, etc. The model helps explain questions the media often poses but fails to answer, such as why politicians never apologize. The paper also shows that economic tools may work well in explaining a social institution that is understudied.

For future research, the theory of dynamic signaling explored here could be generalized. Additionally, the model can be specialized to legal settings and corporate culture or expanded with notions of forgiveness in a larger community of actors. Also, implications for the evolution of emotions such as shame or guilt can be explored. Further empirical tests are possible by looking at court cases, political apologies, corporate scandals or surveys of corporate culture.

6 Appendix A Proof of the Propositions

6.1 Proof of Proposition 1

Proof part (a): Working backward, an agent apologizes if and only if she gets higher payoffs from apologizing:

$$v(b(1,y,p),\theta)-c(\theta,\omega)\geq v(b(0,y,p),\theta)-0 \tag{9}$$

or rearrange to get the apology condition:

$$v(b(1,y,p),\theta)-v(b(0,y,p),\theta)\geq c(\theta,\omega). \tag{10}$$

Since $v(b,\theta)$ is increasing in b, and $c(\theta,\omega)>0$, an apology will only be tendered if $b(1,y,p)>b(0,y,p)$.

Now, let us turn to the principal's beliefs. First note than in Equation (10), the l.h.s. is independent of ω and so for any given θ, there is a $c^*(\theta)$ where the equation is satisfied with equality. Then define:

$$\begin{aligned}\Omega_G^A &\equiv \{\omega : c(\theta^G,\omega)<c^*(\theta^G)\}\\ \Omega_B^A &\equiv \{\omega : c(\theta^B,\omega)<c^*(\theta^B)\}\end{aligned} \tag{11}$$

Note that for some cost functions, these sets may be empty leading to a corner solution where no apologies take place. Again, Proposition 2 gives conditions for when these cost function provide an interior solution.

Also define

$$\begin{aligned}\Omega_G^{\hat{y}} &\equiv \{\omega : y(\theta^G,\omega)=\hat{y}\}\\ \Omega_B^{\hat{y}} &\equiv \{\omega : y(\theta^B,\omega)=\hat{y}\}\end{aligned} \tag{12}$$

The posterior, $b(a,y,p)$, is derived from the prior using Bayes' rule:

$$\begin{aligned} b(1,y,p) &= \Pr[\theta=\theta^G \mid 1,y]\\ &= \frac{F(\Omega_G^1\cap\Omega_G^y)p}{F(\Omega_G^1\cap\Omega_G^y)p+F(\Omega_B^1\cap\Omega_B^y)(1-p)}\\ b(0,y,p) &= \Pr[\theta=\theta^G \mid 0,y]\\ &= \frac{F(\Omega_G^0\cap\Omega_G^y)p}{F(\Omega_G^0\cap\Omega_G^y)p+F(\Omega_B^0\cap\Omega_B^y)(1-p)}\end{aligned} \tag{13}$$

Assuming *b* is not *1* or *0*, we can derive the following from Equation (13)

$$\partial b/\partial p>0 \tag{14}$$

Similarly, by the definition of θ, y is increasing in θ. Thus using a chain rule argument

$$\partial b / \partial y > 0 \tag{15}$$

Finally, since higher θ means both higher y and as I will establish, higher $\Pr[a=1|y]$, higher θ means higher b. If *b* equals *1* or *0*, then the above holds with equality.

Proof part (b): The next result is that good types apologize more than bad types.

Rearranging terms from Equation (13), $b(1,y,p) > b(0,y,p)$ if and only if

$$\begin{aligned} \frac{F(\Omega_B^1 \cap \Omega_B^y)}{F(\Omega_G^1 \cap \Omega_G^y)} &< \frac{F(\Omega_B^0 \cap \Omega_B^y)}{F(\Omega_G^0 \cap \Omega_G^y)} \\ \frac{1-\Pr[a=1|\theta^G, y]}{\Pr[a=1|\theta^G, y]} &< \frac{1-\Pr[a=1|\theta^B, y]}{\Pr[a=1|\theta^B, y]} \end{aligned} \tag{16}$$

This equation holds if and only if

$$\Pr[a=1|\theta^G, y] > \Pr[a=1|\theta^B, y] \tag{17}$$

Good types must apologize more than bad types in order for apologies to be a signal of goodness.

The relationship between the prior, p, and the likelihood of apologies is more complex. As one takes the limit of the principal's beliefs in Equation (13) as p goes to zero or to one, the priors dominate the posteriors:

$$\forall a, y : \lim_{p \to 0,1} b(a,y,p) = p \tag{18}$$

Then by the continuity assumption of $v(b,\theta)$, as the differences in beliefs $b(1,y,p) - b(0,y,p)$ goes to zero, then $v(b(1,y,p)) - v(b(0,y,p))$ goes to zero and thus from the apology condition in Equation (10):

$$\lim_{p \to 0} F(\Omega_\theta^A) = \lim_{p \to 1} F(\Omega_\theta^A) = 0 \tag{19}$$

This implies that when the principal is fairly confident in the agent's type, the prevalence of apologies, $F(\Omega_\theta^A)$, and apology's impact, $b(1,y,p) - b(0,y,p)$ goes to zero. Thus, the prevalence and the impact of an apology is maximized at intermediate values of p when there is uncertainty about the agent's type.

By similar logic, recall Equation (15), $b(a,y,p)$ is increasing in y, but it is bounded by 1. Then so long as $y \in \Re$ has full support and the distribution of $y(\theta^G)$ first order stochastically dominates the distribution of $y(\theta^B)$,[23]

$$\lim_{y \to \infty} b(1,y,p) - b(0,y,p) = 0 \tag{20}$$

For y high enough, the benefit of apologies goes away, and thus unsurprisingly, for good enough outcomes, agents will not apologize. The opposite is also true, for outcomes that are phenomenally bad, apologies also will not help.

[23] This is the only result that requires first order stochastic dominance. and in fact a weaker condition would suffice.

6.2 Proof of Proposition 2

This proposition is simply a restatement of the condition in Equation (17).

6.3 Proof of Proposition 3

Equation (10) shows that for a given $c(\hat{\theta},\hat{\omega})$ the size of the cost is proportional to the shift in the principal's beliefs. Further by construction, higher agent type yields higher principal welfare. Similarly, higher prior is good for the principal if the agent type is good, but bad if the agent type is bad; the principal benefits from accuracy.

To see the effect of changing cost, consider a family of cost functions of the following form:

$$c(\theta,\omega) = \kappa \tilde{c}(\theta,\omega) \tag{21}$$

As $\kappa \to 0$, Equation (10) shows that apologies become cheap so all types apologize with equal frequency until the meaning of an apology—the amount by which an apology shifts beliefs—goes to zero. For κ large, apologies become too costly so neither type apologizes and the apology impact also goes to zero; the existence conditions of Proposition 2 are violated. Since apologies provide a second signal of type without changing the signal value of production output, then if apologies are not in equilibrium, the principal has less information, and his welfare decreases: the principal's welfare is maximized for intermediate costs.

6.4 Proof of Proposition 4

Backward induction allows us to collapse the payouts of each sub-game. Thus all of the parameters of each stage game are identical to the two period game except that we need to reinterpret the payoff function $v(b,\theta)$ as the continuation value. The only restriction we had imposed on $v(b,\theta)$ was that it was increasing in the posterior.

7 Appendix B: Examples of Moral Hazard Games

In the first view, the agent chooses the action x before she learns the realization ω. In this view, x represents the agent's intentions, before the state of the world, ω, is realized. The state might represent new information about the situation or new thinking by the agent. The optimal x given ω would be different than the ex ante optimal x. This difference represents regret. An apology here signals that the outcome resulted from a bad situation rather than bad intentions.

In the second view, x is chosen after both θ and ω are realized. However, only θ is persistent, and thus for future interactions, the principal only cares about θ. In this view, ω represents a temporary mood as in Bernheim and Rangel (2004). The principal cares only about the agent's disposition, θ, but cannot easily tell given the confound of ω. An apology indicates that despite the bad outcome in the past, the principal can expect a good disposition in the future.

In the third view, ω represents the old type, while θ represents new thinking. However, the hidden action, x was chosen before the new thinking was realized. Again, the principal only cares about future interactions, thus an apology signals the change from old to new type.

The necessary common property for these production games is that the agent's type (θ,ω) is translated into an outcome that yields utility $u(\theta,\omega)$ for the agent and utility $y(\theta,\omega)$ for the principal, where $y(\theta,\omega)$ is increasing in θ via some action, x, by the agent. In most examples of interest, outcome is a function both of type and the agent's action: $\tilde{y}(x,\theta,\omega)$. However, in equilibrium the agent's action is uniquely determined by her type, $x^*(\theta,\omega)$; so long as the outcome function $\tilde{y}(x,\theta,\omega)$ has the necessary properties, we can focus on the reduced form:

$$y(\theta,\omega) = \tilde{y}(x^*(\theta,\omega),\theta,\omega)$$

The natural specification of utility is as a function of action, output, and type: $u(x, y, \theta, \omega)$. So long as $\tilde{y}(x,\theta,\omega)$ is invertible in x, it is possible to rewrite $u(x, y, \theta, \omega)$ as $\tilde{u}(y,\theta)$:

$$\tilde{u}(y,\theta) = E_\omega u(x(y,\theta,\omega), y, \theta, \omega)$$

Effectively, instead of choosing an action, the agent is choosing an expected output for the principal. The necessary assumption, then, is a simple restriction on the utility function, that $\tilde{u}(y,\theta)$ has increasing differences in y and θ. Either higher agent types value the principal's utility more, or it is easier for higher agent types to provide principals with higher utility. Thus, by Topkis' Theorem, a higher θ agent maximizing such utility yields higher y.

Recall, however, that the agent's choices are embedded in the larger apology game. Thus, in the larger game, it is necessary for the agent's full utility,

$$U_A(y \mid \theta) = E_\omega\left[\tilde{u}(y,\theta) - c(a,\theta,\omega) + v(b(a, y, p),\theta)\right]$$

to have increasing differences in y and θ. Since the continuation value is composed of the production utility and the cost, a simple sufficient condition is that the production utility is increasing in θ and that the cost function is independent of θ. Alternatively, one could assume that the agent has a sufficiently low discount rate for the future such that the supermodularity of the present is preserved.

This specification for the utility function and production technology is awkward, but it captures many common moral hazard problems. Some examples may help clarify.

The first example is the standard moral hazard with high and low productivity where x represents effort. The agent's cost of effort is increasing in x, but higher types have a lower marginal cost of effort.

$$u(x, y, \theta) = -\frac{x^2}{\theta}$$
$$y(x,\omega) = x + \omega$$

Assuming the noise term, ω has mean zero, this expression yields

$$\tilde{u}(y,\theta) = -\frac{y^2 + Var(\omega)}{\theta}$$

which can be differentiated to show that it satisfies increasing differences.

A second example is a political game, with a uni-dimensional policy space, and x represents the agent's choice of policy. The principal has an ideal point of zero, while the agents have ideal points away from zero, and $1/\theta$ represents the agent's ideal point. This functional form can also model the Paul and Amy interaction, where x is the choice of departure time, and ω is the amount of traffic:

$$u(x, y, \theta) = -(x + \omega - \frac{1}{\theta})^2$$
$$y(x,\omega) = -(x + \omega)^2$$

A third example might have θ as an altruism parameter, and the task is some noisy gift giving game, where the agent's choice x is how much the agent gives to the principal, but the agent's choice is obscured by noise ω.[1]

$$u(x, y, \theta) = \theta y - x$$
$$y(x,\omega) = x + \omega$$

The point again of ensuring supermodularity is to guarantee that higher types, θ, make choices that lead in expectation to higher utility for the principal, y. Having established this requirement, I return to the reduced form specification.

8 Appendix C –Cheap Talk Models of Apologies

8.1 Contracting with Commitment: Social Contracts "I'm sorry I'll never do it again"

One approach to modeling cheap apologies is to take a contracting approach where the principal can commit to future termination strategies, creating a mechanism for ensuring truthful revelation. In the other sections, Markov perfection limits the principal to choosing the most attractive agent. Here, I proceed with the same model from Section 2, except here there is no explicit cost of apologies, and I use the solution concept of a Markov Perfect Equilibrium that uses both the principal's beliefs and the agent's apology in the prior period as the state variable. I also allow the principal to commit ex ante to a mixed strategy of retaining the agent or not as a function of agent's apology and outcome. I then look for a separating equilibrium where good types always apologize and bad types never do.

Once again, agents produce an output and then choose whether to apologize or not. Principals now, instead of merely choosing between {continue, terminate}, now can choose $\delta_t(a_{t-1}, y_t)$, the probability of termination, as a function of the apology in the previous period, and the current period outcome.[24]

In a separating equilibrium, beliefs as a function of apologies would be

$$b(1, y) = 1 \qquad b(0, y) = 0 \tag{22}$$

To implement such an equilibrium, the principal chooses $\delta(a, y)$ to arrive at a payoff function, $v(b, \theta)$, such that the following incentive compatibility constraints are satisfied:

$$v(1, \theta^G) \geq v(0, \theta^G) \qquad v(0, \theta^B) \geq v(1, \theta^B) \tag{23}$$

One way to obtain such a payoff function in the first stage of a two-period game is to have appropriate payoffs in the second stage. I call this a "fool me once, shame on you, fool me twice, shame on me" contract. The intuition is that the principal would like to know the private information of the agent, but the agent has incentive to misrepresent her type; an apology is a claim to be a good type. However, if the agent claims she is a good type, the principal will demand much more out of the agent, and tolerate failure far less, whereas if the agent does not apologize, the principal will be more forgiving of failure.

To simplify the problem, assume there are only two possible outputs for the principal so that $y \in \{\underline{y}, \bar{y}\}$.[25] Define the following:

$$s_G = \Pr[y(\theta^G, \omega) = \bar{y}] = F(\{\omega : y(\theta^G, \omega) = \bar{y}\}) \qquad s_B = \Pr[y(\theta^B, \omega) = \bar{y}] = F(\{\omega : y(\theta^B, \omega) = \bar{y}\}) \tag{24}$$

The utility of the agent is given by:

$$U(a, \theta) = u(\theta, \omega_1) + u(\theta, \omega_2) + \delta(a_1, y(\theta, \omega_2))v(\theta) \tag{25}$$

Then the agent's IC constraints so that only good types apologize are

$$E_\omega[\delta(1, y(\theta^G, \omega_2))v(\theta^G)] \geq E_\omega[\delta(0, y(\theta^G, \omega_2))v(\theta^G)] \qquad E_\omega[\delta(0, y(\theta^B, \omega_2))v(\theta^B)] \geq E_\omega[\delta(1, y(\theta^B, \omega_2))v(\theta^B)] \tag{26}$$

which can be simplified using Equation (24) and rearranged to yield

[24] It would also be sensible to allow δ_t to depend on a_t, but I do not to simplify the analysis.
[25] These results generalize easily to continuous θ. They are messier but similar for continuous y.

(27) $$\frac{s_G}{1-s_G} \geq \frac{\delta(0,\underline{y}) - \delta(1,\underline{y})}{\delta(1,\bar{y}) - \delta(0,\bar{y})} \geq \frac{s_B}{1-s_B}$$

Moral hazard concerns provides two more constraints—an agent in the second round must get higher payoffs for success than failure:

(28) $$\delta(1,\bar{y}) > \delta(1,\underline{y})$$
$$\delta(0,\bar{y}) > \delta(0,\underline{y})$$

Combining these constraints gives us the following ordering of $\delta(a,y)$:

(29) $$\delta(1,\bar{y}) > \delta(0,\bar{y}) > \delta(0,\underline{y}) > \delta(1,\underline{y})$$

Effectively, the marginal benefit of success in the second stage in the case of an apology in the first, must be higher than the marginal benefit in case of no apology in the first. An apology will lead the principal to believe the agent is a good type, but he will expect better performance from her in the future. In the unreduced form, this means higher effort for the agent after an apology.

Alternatively, the following ordering is also possible given the constraints:

(30) $$\delta(0,\bar{y}) > \delta(1,\bar{y}) > \delta(1,\underline{y}) > \delta(0,\underline{y})$$

I argue this ordering is unlikely by considering the principal's maximization problem. If the probability of the bad type succeeding is still relatively high, which is likely if the principal has difficulty differentiating between bad and good, then the principal would prefer the ordering given in Equation (29).

Given the above analysis, I now return to the case of Paul and Amy. Paul may tolerate one failure from Amy to show up on time, but given the repeated failure, he is forced to end the relationship.

This game could be extended to N-periods except that after the second period, there is complete separation making signaling uninteresting. To get around this, I relax the complete persistence of type. Let there be some probability of mutation:

(31) $$\Pr[\theta_{t+1} = \theta^G \mid \theta_t = \theta^G] = \bar{p}$$
$$\Pr[\theta_{t+1} = \theta^B \mid \theta_t = \theta^B] = \underline{p}$$

Then even though there is full separation each period, the two period equilibrium can also serve as a Markov perfect equilibrium for an n-period game where the prior is reset to either $\bar{p}$ or $\underline{p}$ after each period.

In any case, the problem with the contract presented in this section is that it is not renegotiation proof. Once an agent has apologized, she has established herself as the good type. At that point, the principal would not want to end the relationship. To solve this problem I introduce the ability for the principal to offer different tasks.

8.2 Contracting without Commitment: Status "I'm sorry; I'm an idiot."

8.2.1 Introducing tasks

One reason why contracting is difficult for the principal in the previous case is the limitation of the principal's strategy space. In this section, I give apologies more sophisticated meaning by expanding the space of principal responses. Now, instead of "continue" or "terminate," I allow the principal to offer a menu of tasks. Let there be a set Z of tasks for each period, each task defined by an ordered triple $(\delta_z, \rho_z, \phi_z)$, where the discount rate δ_z reflects how long before that task comes up again and ρ_z is the correlation of the next task with the current task in the θ dimension and ϕ_z is the correlation in the ω dimension. Until now, I have assumed that θ is identical across periods, while ω was drawn independently, limiting the principal's choice set for the next period to

(32) $$Z = \{(\delta_{cont} = 1, \rho_{cont} = 1, \phi_{cont} = 0), (\delta_{term} = 0, \rho_{term} = 0, \phi_{term} = 0)\}$$

In this section, I expand the set of tasks available to the principal to a larger set. As before, θ represents an internal dimension or disposition, while ω represents an external dimension or situation, but here I consider

scenarios where the agent is expected to have some control over her situation. Whereas before, an agent could excuse poor performance to a bad situation, doing so now would admit to a lack of control, which is also bad for the agent. Formally, the change in the model is that now, payoffs to the agent are based not just on the principal's beliefs about θ, but also the principal's beliefs about ω as well.

Returning to the example of Paul and Amy, let the base task be "be on time." Let θ represent how much Amy cares about Paul, and let ω represent how able Amy is at showing up to events on time. Now, consider two other tasks that Paul might like fulfilled: "talk to at party," and "be on time for job interview." One would expect that success at "talk to at party" would be correlated with how much Amy cares about Paul, but not be correlated with how good Amy is at showing up on time for things. Conversely, "be on time for job interview" might depend very much on Amy's ability to show up on time for things, but not depend on Amy's liking of Paul. Thus one might imagine an equilibrium in which if Amy is late and apologizes, then Paul would be happy to talk to Amy if he sees her at a party, whereas if she did not apologize, Paul might be more likely to recommend Amy for a job opening that depended on her on time arrival for the interview.

This expansion of the set of tasks available allows the principal to offer a renegotiation proof menu that allows cheap apologies to carry meaning even if the principal is not able to commit to a contract.

8.3 Model Details

The setup of the game starts again with the base model, and once again, the cost of apology is set to zero. The main change is that now, both θ and ω are semi-persistent, and the degree of persistence—i.e. correlation, across periods—is determined by the principal's choice of tasks. The agent's choice of actions each period is the same as before, she produces an output for the principal, and then, upon realization of the output, decides to apologize or not. The principal offers a menu of two future tasks, one if the agent apologizes, and the other if she does not. The choice of task determines the distribution that governs the agent's type $(\theta_{t+1}, \omega_{t+1})$ in the next period.

The principal's payoffs are the same: maximize his output across periods.

(33) $$U_P = \sum_t y(\theta_t, \omega_t)$$

The agent's per period payoff is also the same, except now, the agent's continuation value depends not just on the principal's beliefs about her internal type, b_θ, but also the principal's beliefs about her external type, b_ω:

(34) $$U_A = u(\theta_t, \omega_t) + v(b_\theta, \theta_t, b_\omega, \omega_t)$$

In a two period game, the agent's second period payoff will depend on the task selected by the principal at the end of the first.

(35) $$U_A = u(\theta_t, \omega_t) + \delta(z_{t+1}(a_t, y_t))u(\theta_{t+1}, \omega_{t+1})$$

I add a number of simplifying assumptions. None are crucial, but they make analysis more comprehensible. Limit the external dimension to two values so that $\omega \in \{\omega^L, \omega^H\}$ one of which represents low ability, and the other representing high ability. I retain $\theta \in \{\theta^G, \theta^B\}$ so that there are good agents who care about the principal, and bad agents who do not care. Then, for any particular task in period *t*, the agent's type is given by (θ_t, ω_t) which can take one of four values.

Now, assume also that there are only two possible outputs for the principal, $y \in \{0,1\}$, a task can be a failure or a success. Assume also that an agent succeeds at a given task only if she is both of good disposition, and high ability: (θ^G, ω^H):

$$
\begin{aligned}
y(\theta^G, \omega^H) &= 1 \\
y(\theta^G, \omega^L) &= 0 \\
y(\theta^B, \omega^H) &= 0 \\
y(\theta^B, \omega^L) &= 0
\end{aligned}
\tag{36}
$$

Similarly, assume that agent's utility is increasing in type. Specifically, assume the agent's utility for consumption is given as:

$$
\begin{aligned}
u(\theta^G, \omega^H) &= 1 \\
u(\theta^G, \omega^L) &= 0 \\
u(\theta^B, \omega^H) &= 0 \\
u(\theta^B, \omega^L) &= 0
\end{aligned}
\tag{37}
$$

Note that the agent is also only happy when she is successful. In this simplified form, there is preference alignment between principal and agent, when it comes to the task at hand. This alignment is not necessary, but it makes the conflict of interest introduced by task assignment more apparent.[26]

Some notation will be helpful. Recall for a given task z to be assigned in the next period, ρ_z is the correlation of the new θ_{t+1} with the current θ_t, while ϕ_z is the correlation of ω_{t+1} with ω_t. The principal's prior that the agent is θ^G for a given task z is p_z and the prior that the agent is ω^H for a given task z is q_z. Define:

$$
\begin{aligned}
\overline{p}_z &= \Pr[\theta_{t+1} = \theta^G \mid \theta_t = \theta^G, z] = p_z + \tfrac{\rho_z}{p}\sqrt{p(1-p)p_z(1-p_z)} \\
\underline{p}_z &= \Pr[\theta_{t+1} = \theta^G \mid \theta_t = \theta^B, z] = p_z - \tfrac{\rho_z}{1-p}\sqrt{p(1-p)p_z(1-p_z)} \\
\overline{q}_z &= \Pr[\omega_{t+1} = \omega^H \mid \omega_t = \omega^H, z] = q_z + \tfrac{\phi_z}{q}\sqrt{q(1-q)q_z(1-q_z)} \\
\underline{q}_z &= \Pr[\omega_{t+1} = \omega^H \mid \omega_t = \omega^L, z] = q_z - \tfrac{\phi_z}{1-q}\sqrt{q(1-q)q_z(1-q_z)}
\end{aligned}
\tag{38}
$$

I now look for a renegotiation-proof Markov Perfect Equilibrium in pure strategies again using beliefs and last period apologies as a state space. Given the stark production technology specified in Equation (36), if the principal observes a success, $y = 1$ then he knows for sure that the agent is (θ^G, ω^H) and seeks to assign a task as similar to the current task as soon as possible. That is, a task where δ, ρ, ϕ are all close to one. If Amy shows up on time, then Paul will ask her back to meet again the following week.

In the event of a failure, $y = 0$, the set of possible agent types narrows to $\{(\theta^G, \omega^L), (\theta^B, \omega^H), (\theta^B, \omega^L)\}$. Assuming that success is relatively common so that both p and q are relatively high, the third case would be rare. Consider a strategy by the principal that allows him to distinguish between the first two cases.

The principal offers a menu of two tasks after a failure. An agent who apologizes gets task z_1, and an agent who does not apologize gets task z_0. To get separation, the principal selects tasks so that (θ^G, ω^L) types apologize and (θ^B, ω^H) types do not. The incentive compatibility conditions are:

[26] As before, one could introduce moral hazard via a hidden action for the agent that would yield the same reduced form.

$$\delta(z_1)\bar{p}_{z1}\underline{q}_{z1} \geq \delta(z_0)\bar{p}_{z0}\underline{q}_{z0}$$
$$\delta(z_0)\underline{p}_{z0}\bar{q}_{z0} \geq \delta(z_1)\underline{p}_{z1}\bar{q}_{z1} \quad (39)$$

Ideally, the principal would offer one task perfectly correlated on the internal dimension, $(\delta_{z1}=1,\rho_{z1}=1,\phi_{z1}=0)$, and the other task perfectly correlated on the external dimension, $(\delta_{z0}=1,\rho_{z0}=0,\phi_{z0}=1)$, but the actual assignment depends upon task availability. The conflict arises because payoffs for the principal are undiscounted. I assume that the principal interacts with potentially many agents, and thus could have the task filled by another. Thus, the difference in correlations must be sufficient to overcome the difference in discount rates. For example, in the example with Paul and Amy, the "job interview" task may come quite infrequently, and would have a particularly low discount rate. If no appropriate task is available, then apologies would be uninformative. Often, agents apologize consequence free.

In the case of politics, as demonstrated by Lee and Tiedens (2001), an apology gains favor in the "liking" domain at the cost of the "respect" domain. A president who apologizes for a sexual indiscretion might be liked more, and thus given taks based on liking, such as "dating my daughter," but would not be given further tasks based on judgment, such as "running the country."

Returning to the aforementioned gender and cultural differences, it may be more difficult for men to apologize because they encounter more often (higher δ's) tasks based on status or competence. An evolutionary reason might be because women use measures of status and competence to choose their mates (Cole, Mailath and Postlewaite, 2001). In terms of culture, one might find that cultures that apologize more, such as in Japan, have production technologies based on group production, where preference alignment is more valuable, while in Western cultures, performance pay tied to individual competence is more common.

8.4 Partial Apologies: Empathy "I'm sorry your grandmother died"

In this section, I consider another possible dimension of type, but instead of control, I consider *empathy*, a measure reflected in the game's information structure.

Though the primary purpose of an apology—and the primary dictionary definition—is in relation to a fault or offense, the notion of apologies is often conflated with a general sense of empathy, or awareness of the other's emotional state: e.g. "I am sorry to hear that your grandmother died." Alternatively, empathy can be interpreted as awareness by the agent of what the principal considers appropriate rules of conduct. This section presents a model of partial apologies: those apologies that do not come with an admission of guilt.

To capture this interaction, return to the base model where again, an apology will be a shift in attribution from preference alignment, θ, to environment, ω. Now, let type be three dimensional instead of two, given by the triple (θ,ω,τ) for each period. In addition to a preference alignment type, θ, let there be an empathy type, $\tau\in\{0,1\}$, where empathic and non-empathic types differ in their information sets; non-empathic types do not observe the principal's payoff, y.[27]

Let there be a positive correlation, ψ, between θ and τ, either because the empathic types are more effective at producing given their better understanding of the principal, or for some external reason such as common upbringing. Let the prior on τ be defined as $q=\Pr[\tau=1]>\frac{1}{2}$.

Again, there is no cost of apology so the agent receives utility only from production. Also, this model again restricts the principal's choice set to either stay with the current agent or switch to the outside option. The probability that the current agent is better than the outside option is again given by $\delta(b_\theta(a,y))$. In a two period game, the agent's utility is

$$U_A(a\mid\theta,\omega_1,\omega_2,\tau)=u(\theta,\omega_1)+\delta(b_\theta(a,y))u(\theta,\omega_2) \quad (40)$$

[27] Empathy could be made continuous by specifying information sets over states of the world, ω, rather than outcomes, y, with agents that have greater empathy having a finer partition. However, I again favor simplicity.

Assume that output, $y(\theta,\omega)\in\{0,1\}$, takes only two values, success and failure.

In such a game with cheap apologies, the agent's strategy is given by her apology decision $a\in\{0,1\}$. This decision is conditioned on y for empathic types, but non-empathic types have only one information set, and so their only pure strategy would be "always apology" or "never." Consider the following equilibrium: Empathic types always apologize in case of failure, and never apologize in case of success. Non-empathic types never apologize.

In this equilibrium, the principal's beliefs about the agent's empathy, τ, is

$$\begin{aligned}&\Pr[\tau=1\mid a=0,y=0]=0\\&\Pr[\tau=1\mid a=1,y=0]=1\\&\Pr[\tau=1\mid a=0,y=1]=q\\&\Pr[\tau=1\mid a=1,y=1]=0\end{aligned}\tag{41}$$

An appropriate apology proves empathy, and conveys information about θ via its positive correlation with τ. Success is never accompanied by an apology and thus provides no information. If success is followed by an apology, this is off the equilibrium path, and I specify that this indicates a non-empathic type. An inappropriate apology automatically reveals the non-empathic types.

The principal's updated beliefs regarding the agent's θ, are

$$b_\theta(a=0,y=1)=\frac{F(\Omega^G)p}{F(\Omega^G)p+F(\Omega^B)(1-p)}$$

$$b_\theta(a=1,y=0)=\frac{F(\Omega\setminus\Omega^G)\Pr[\tau=1\mid\theta=\theta^G]p}{F(\Omega\setminus\Omega^G)\Pr[\tau=1\mid\theta=\theta^G]p+F(\Omega\setminus\Omega^B)\Pr[\tau=1\mid\theta=\theta^B](1-p)}\tag{42}$$

$$b_\theta(a=0,y=0)=\frac{F(\Omega\setminus\Omega^G)\Pr[\tau=0\mid\theta=\theta^G]p}{F(\Omega\setminus\Omega^G)\Pr[\tau=0\mid\theta=\theta^G]p+F(\Omega\setminus\Omega^B)\Pr[\tau=0\mid\theta=\theta^B](1-p)}$$

Using the positive correlation between τ and θ yields

$$\begin{aligned}&\Pr[\tau=1\mid\theta=\theta^G]>q>\Pr[\tau=0\mid\theta=\theta^G]\\&\Pr[\tau=0\mid\theta=\theta^B]>q>\Pr[\tau=1\mid\theta=\theta^B]\end{aligned}\tag{43}$$

so we can get:

$$b_\theta(a=1,y=0)>b_\theta(a=0,y=0)\tag{44}$$

Empathic types follow this equilibrium, because doing so increases the principal's beliefs that the agent is a high type, *b(a,y)*, and hence increases payoffs. Deviations would decrease payoffs. Non-empathic types have no information about the state of the world and the payoffs to the principal. Since non-empathic agents cannot condition their apologies on the state of the world, they must choose to either always apologize or always not apologize. The equilibrium strategy is optimal for the non-empathic types so long as the expected benefit of not apologizing is greater than the expected benefit of always apologizing:

$$\begin{aligned}&F(\Omega^i)\delta(b_\theta(0,1))+F(\Omega\setminus\Omega^i)\delta(b_\theta(0,0))\\&\quad>F(\Omega^i)\delta(b_\theta`(1,1))+F(\Omega\setminus\Omega^i)\delta(b_\theta(1,0))\end{aligned}\tag{45}$$

Or re-arrange to get the probability of success times the marginal benefit of apologizing in case of success must be greater than the probability of failure times the marginal benefit of apologizing in case of failure:

$$F(\Omega_B^1)[v(b(0,1)) - v(b(1,1))] > F(\Omega \setminus \Omega_B^1)[v(b(1,0)) - v(b(0,0))] \tag{46}$$

Essentially, if in most situations, the agent is successful ($F(\Omega_B^1) >> F(\Omega \setminus \Omega_B^1)$) and an apology is typically unwarranted, then the non-empathic agent finds it optimal to never apologize.

It is useful to note that once the apology is made, in a simplified model without any further noise and type is stable across time, the principal learns for sure that the agent is empathic. Thus over repeated interactions, the principal quickly becomes aware that an agent is empathic. However, repeated failures would still lead the principal to conclude the agent is a bad type, and thus the principal will end the relationship anyway. This rapid devaluation of the perfunctory apology corresponds to the real world observation that often such apologies seem meaningless. Once empathy has been established, further apologies have little impact on the principal's beliefs regarding the agent's type. The relatively minor impact of apologies in this scenario accords with the assertion that these apologies—apologies without admission of fault—are only partial apologies. However, if ever an agent fails to offer a partial apology when it is expected, judgments can shift quickly against her.

Another interpretation of this model is in the situation where an apology is tendered before the principal is even aware of the mistake. An apology demonstrates awareness than a transgression occurred and that an apology is warranted.

Alternatively, one might think of a world where different standards of behavior are possible. In one culture, being an hour late is unacceptable while for another, being an hour late is a virtue. An apology can be thought of as an acknowledgement by the agent that she violated a norm according to the standards of the principal. An apology indicates a shared agreement of the norms of behavior, or at the very least, an awareness by the agent of what the principal considers are the norms of behavior.

In the example of Paul and Amy, the first apology for her tardiness demonstrated to Paul that Amy knows enough to apologize for her mistakes, that she is aware of Paul's feelings, or that she acknowledges Paul's view that the tardiness is a mistake. However, once empathy has been established, repeated apologies no longer help. After the third time, Paul effectively concludes that Amy may be aware of his feelings but she is still of a bad type.

Incidentally, this empathy variant applies equally well for other perfunctory pleasantries such as "thank you" or "congratulations."

9 Appendix D – Experiment Instructions

Instructions Thank you for participating in this study of economic relationships in the presence of interaction.

Please be advised that there is no talking once the experiment has begun, except to ask questions. I will be available to answer questions especially during the practice game. Also, kindly turn off all cell phones.

Note: This is a study for an economics research project. It is the norm of the experimental economics profession for the experimenter to never deceive subjects. Please be assured that the game will proceed exactly as described here.

The Experiment There is a number written in the top right corner of this page. That is your ID number. It will be necessary so that proper payouts can be calculated.

You have been randomly assigned to one of the two experiment rooms. If you are assigned to room L5, you will be given the role of First Mover. If you are assigned to room L8, you will be given the role of Second Mover.

This experiment is designed to study two person relationships. You will play the same game five times, each time with a new partner. In each game, each First Mover will be paired with a randomly selected Second Mover to play a game that will last 10 periods. After each game is over, the Second Movers will switch seats and be paired with a new First Mover

The Game Each game will last for 10 periods. You will be interacting with the same partner throughout the 10 periods. You begin each game with 120 tokens. You can earn more tokens throughout the course of the game. At the beginning of each game, a new communication cost will be selected to be used for the duration of the game.

In each period, three things happen.

1) The First Mover (FM) is given 10 tokens to allocate. These tokens can be used for one of two purposes. Each token can either be allocated to the Second Mover (SM), or the tokens can be banked. The FM can allocate between 1-10 tokens to the SM. Tokens not allocated are banked.

2) The SM receives a number of tokens equal to the number the FM allocated **times three**. The SM can also do one of two things, keep the tokens, or allocate the tokens to a project that benefits only the FM. If the project is successful, the FM receives 20 tokens in addition to the ones he previously banked. If the project is a failure the FM receives 0 tokens in addition to the ones he previously banked. Each additional token that the SM allocates to the FM's project increases the probability of success by 5%.

The SM's earnings comes only from tokens he does not allocate to the FM's project. The FM's earnings comes both from tokens that he banked and from tokens earned from successful projects.

3) Both the FM and the SM observe the outcome of the project. The FM observes only whether the project was successful, but **does not observe** the SM's allocation. The SM now can choose whether to pay the communication cost and send a simple message that reads, "I am sorry."

The period thus ends and repeats to step 1.

In Summary, the payoffs for each period are

FM's tokens = 10 – Number Entrusted + Project Earnings

SM's tokens = Number Banked – Communication Costs

Payment At the end of each game, the profits will be displayed. Each SM will switch seats and be paired with a new FM, and a new game will begin with profits reset to 120 tokens. After all the games are complete, one of the games will randomly (by six sided die) be selected for payment. You are only paid the profits that you earn for the randomly selected game. However, it is equally likely for any particular game to be selected, so try to earn as much as possible in each game.

The exchange rate for tokens to dollars: 120 tokens is worth $10. You will be paid rounded up to the nearest dollar.

If you have any questions please ask them at this time.

There will be a practice round before we begin.

10 References

Aaker, Jennifer, Susan Fournier. and Adam Brasel. 2004. "When good brands go bad." *Journal of Consumer Behavior*, 31:1-16.

Abeler, Johannes, Calaki, Juljana, Andree, Kai and Basek, Christoph, (2009), The Power of Apology, No 2009-12, Discussion Papers, The Centre for Decision Research and Experimental Economics, School of Economics, University of Nottingham.

American Heritage Dictionary, Fourth Edition (2000)

"apology" (2008). In *Merriam-Webster Online Dictionary*. Retrieved June 20, 2008.

Becker, Gary. 1976. *The Economic Approach to Human Behavior*. Chicago: University of Chicago Press.

Bennett, Mark and Carla Dewberry. 1994 "I've Said I'm Sorry, Haven't I - A Study Of The Identity Implications And Constraints That Apologies Create For Their Recipients," *Current Psychology*, 13(1):10-20.

Bennett, Mark and Deborah Earwaker. 2001 "Victim's Responses to Apologies: The Effects of Offender Responsibility." *Journal of Social Psychology.* 134:457-464.

Berg, Joyce, John Dickhaut and Kevin McCabe. 1995. "Trust Reciprocity and Social History" *Games and Economic Behavior* 10(1):122.

Bernheim, Douglas, and Antonio Rangel. 2004 "Addiction and Cue-Triggered Decision Processes" *American Economic Review*. 94(5): 1558-1590.

Camerer, Colin. 1988. "Gifts as Economic Signals in Social Systems." *Amer Journal of Sociology*. 94:S180-S214.

Charness, Gary. and Martin Dufwenberg. 2006. "Promises and Partnerships." *Econometrica*. 74(6):1579-1601.

Chaudhuri, Ananish. and Lata Gangadharan. 2007 "An Experimental Analysis of Trust and Trustworthiness ", Southern Economic Journal, 73 (4): pp. 959 - 985

Chaudhuri, Ananish. and Lata Gangadharan. 2003. "Sending Money in the Trust Game: Trust or Other-Regarding Preferences?" http://papers.ssrn.com/sol3/papers.cfm?abstract_id=479961

Cole, Harole., George. Mailath, and Andrew Postlewaite 2001. "Efficient Non-Contractible Investments in Large Economies," *Journal of Economic Theory.* 101:333-373.

Cox, James. 2002. "Trust, Reciprocity and Other-Regarding Preferences: Groups vs. Individuals and Males vs. Females", Chapter 14 in Rami Zwick and Amnon 27 Rapoport (Eds.), *Experimental Business Research*, Kluwer Academic Publishers, Boston, Dordrecht, London.

Ekman, Paul. 1969. "Nonverbal leakage and clues to deception." *Psychiatry.* 32(1): 88-106.

Ensminger, Jean. 2000, "Experimental Economics in the Bush: Why Institutions Matter", in Menard, C., ed., *Institutions, Contracts and Organizations*, Northampton, MA: Edward Elgar.

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

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

Frank, Robert. 1989. *Passion Within Reason*, W.W. Norton and Company.

Frank, Robert .2004. "In Defense of Sincerity Detection." *Rationality and Society*. 16(3) 287-305.

Friedman, Milton. 1953. *Essays in Positive Economics*. Chicago: U. of Chicago Press.

Gneezy, Uri., Werner Güth, and Frank Verboven. 2000, "Presents or Investment? An Experimental Analysis." *Journal of Economic Psycholog*. 21: 481-93.

Ho, Benjamin and Elaine Liu 2009. "Apologies in Medical Malpractice: The impact of I'm Sorry Laws" mimeo.

Heider, Fritz. 1958. The psychology of interpersonal relations, New York: Wiley

Jones, Edward and Richard Nisbett. 1972. "The actor and the observer: Divergent perceptions of the causes of the behavior." In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins and B. Weiner (eds.), Attribution: Perceiving the causees of behavior (pp. 79-94). Morristown, NJ: General Learning Press.

Kreps, David, Paul Milgrom, John Roberts, and Robert Wilson. 1982 "Rational Cooperation in the Finitely Repeated Prisoners Dilemma'." *Journal of Economic Theory*. 27: 245-252.

Koford, Kenneth, 1995. "Trust and Reciprocity in Bulgaria: A Replication of Berg, Dickhaut and McCabe" University of Delaware, Department of Economics Working Paper: 1998/08.

Latif, E. 2001. "Apologetic Justice: Evaluating Apologies Tailored Toward Legal Solutions" 81 B.U.L. Rev. 289.

Lee, Fiona, Christopher Peterson, and Larissa Tiedens. 2004. "Mea Culpa: Predicting Stock Prices from Organizational Attributions." *Personality and Social Psychology Bulletin*. 30(12):1636-1649.

Lee, Fiona and Larissa Tiedens. 2001a. "Who's being served? "Self"-serving attributions and their implications for power." *Organizational Behavior and Human Decision Processes*, 84(2), 254-287.

Lee, Fiona and Larissa Tiedens.2001b. "Is it lonely at the top? Independence and interdependence of power-holders." In B. Staw and R. Sutton (Eds.), Research in Organizational Behavior, Vol. 23, p. 43-91.

Maskin, Eric and Jean Tirole. 1990 "The Principal-Agent Relationship with an Informed Principal: The Case of Private Values *Econometrica*. 58 (2): 379-409.

McCullough, M., Rachal, K., Sandage, S., Worthington, E., Brown, S, Hight, T. 1997. Interpersonal Forgiving in Close Relationships: II. Theoretical Elaboration and Measurement. *Journal of Personality and Social Psychology*. 75(6) p1586-1603.

Ohbuchi, Ken-ichi, Kameda, Masuyo. , Agarie, Nariyuki. 1989. "Apology as Aggression Control: Its Role in Mediating Appraisal of and Response to Harm." *Journal of Personality and Social Psychology*. 59(2): 219-227.

Prendergast, Canice and Lars. Stole. 2001. "The non-monetary nature of gifts.' *European Economic Review* 45(10) p 1793-1810.

Ross, Lee. 1977. "The intuitive psychologist and his shortcomings: Distortions in the attribution process." In L. Berkowitz (ed.), *Advances in experimental social psychology* (Volume 10, pp. 173-240), Orlando, FL: Academic Press

Sally, D. 2001. "On Sympathy and Games" Journal of Economic Behavior and Organization.

Sally, D. 2002 "Two economic applications of sympathy" Journal of Law Economics and Organization.

Schweitzer Maurice, Eric Bradlow and John Hershey. 2006. 'Promises and Lies: Restoring Violated Trust." Organizational Behavior and Human Decision Making..

Spence, Michael. 1974. "Competitive and Optimal Responses to Signals: An Analysis of Efficiency and Distribution." *Journal of Economic Theory*, 7(3):296-332.

Takaku, Seiji., Weiner, Bernard. and Ohbuchi, Ken-ichi. 2001. "A Cross-Cultural Examination of the Effects of Apology and Perspective Taking on Forgiveness" *Journal of Language and Social Psychology*, 20(1):144-166.

Tavuchis, Nicholas. 1991. *Mea Culpa: A Sociology of Apology and Reconciliation*, Stanford, CA: Stanford University Press.

Tiedens, Larissa. 2001. "Anger and advancement versus sadness and subjugation: The effect of negative emotion expressions on social status conferral," *Journal of Personality and Social Psychology*. 80 (1): 86–94.

Weiner, Bernard., Graham, S., Peter, O., Zmuidinas, M. 1991. "Public Confession and Forgiveness." *Journal of Personality*, 59(2)

Xiao, Erte and Daniel Houser. 2005. "Emotion Expression in Human Punishment Behavior." *Proceedings of the National Academiy of Science*. 102(20): 7398-7401.

11 Figures

Figure 1

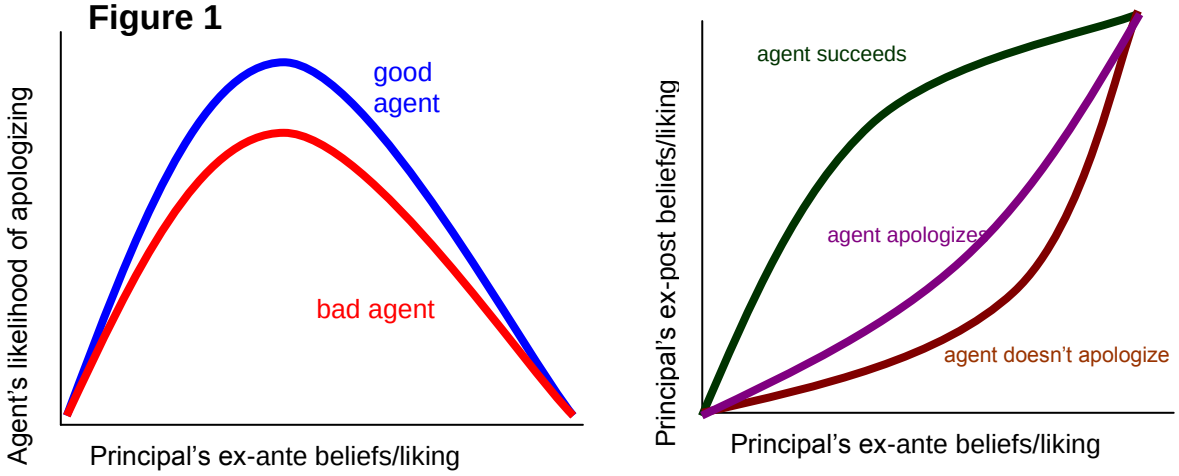
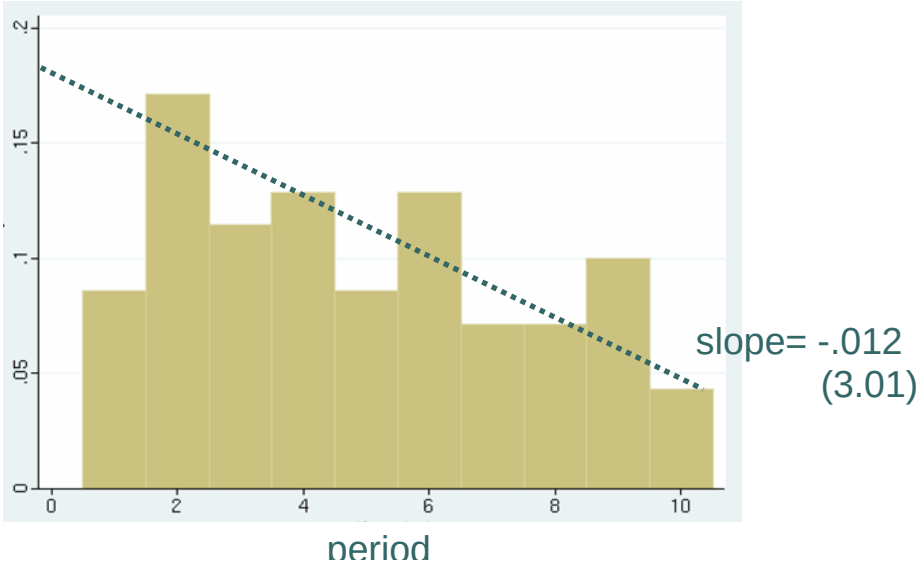


Figure 2



Frequency an apology is given if project failed, by period. (t-stat in parentheses)

12 Tables

Table 1: Summary Statistics

Summary Statistics by Communication Cost (stdev in parentheses)

	c=0 n=330	c=5 n=270	c=10 n=350	c=15 n=300
% who apologize	38	15	6.2	2.7
tokens entrusted	5.31	5.29	5.31	5.55
	(3.84)	(3.63)	(3.71)	(3.64)
tokens returned	5.76	4.8	5.48	5.45
	(6.03)	(5.49)	(6.13)	(5.79)
percent return	29	26	28	29
	(30)	(29)	(29)	(28)
principal profits	10.4	10.3	10.3	9.92
	(7.89)	(8.03)	(8.03)	(7.99)
agent profit	10.2	10.3	9.8	10.8
	(8.37)	(8.17)	(8.45)	(8.58)

Table 2

Amount entrusted in periods 1-9

	Entrust at t+1 (b)	Entrust at t+1 (b)	Entrust at t+1 (b)
Apologize (a)	1.61	1.49	0.74
	(0.30)**	(0.30)**	(0.28)**
Project Earnings (y)	0.24	0.22	0.14
	(0.01)**	(0.01)**	(0.01)**
Avg Percent Returned (θ)		0.14	0.07
		(0.04)**	(0.03)*
Entrust at t (p)			0.44
			(0.03)**
const	3.44	2.78	1.32
	(0.13)**	(0.20)**	(0.20)**
# of Obs	1125	1125	1125
R^2	0.32	0.33	0.46

** p < .01, * p <.05, (White std. err. in parentheses)
Period fixed effect included in all regressions

Table 3

Apology and project earnings in each period:

	Apologize (a)		Project Earnings (y)	
Avg Percent Returned (θ)	0.013		0.79	
	(0.004)**		(0.08)**	
Entrust at t (p)	0.034	0.034	0.88	0.91
	(0.014)**	(0.014)**	(0.064)**	(0.066)**
[Entrust at t (p)]^2	-0.0014	-0.0028		
	(0.0012)	(0.0013)*		
Project Earnings (y)	-0.016			
	(0.0012)**			
const	0.045		-5.96	
	(0.030)		(0.82)**	
Agent Fixed Effect	No	Yes	No	Yes
# of Obs	1250	1250	1250	1250
R^2	0.12	0.007	0.28	0.22

** p < .01, * p <.05, (White std. err. in parentheses)
Period fixed effects included in all regressions.

Table 4

Effect of an apology on continuation values.

	Continuation Value v(-)		
Tokens spent on apology c(-)	1.37	0.97	0.69
	(0.35)**	(0.33)**	(0.33)*
Entrust at t (p)		2.33	1.80
		(0.22)**	(0.22)**
Constant	0.06	7.08	6.09
	(2.13)	(2.14)**	(2.03)**
Agent Fixed Effects	No	No	Yes
Observations	900	900	900
R-squared	0.50	0.56	0.58

** p < .01, * p <.05, (White std. err. in parentheses)

Table 5

Effect of alternate messages on future trust

	0.18	0.13
Project Earnings (y)	(0.02)	(0.02)
	1.24	0.38
"I am happy"	(0.63)	(0.58)
	1.01	0.58
"I am sad"	(0.76)	(0.69)
	2.42	1.40
"I am sorry"	(0.79)	(0.73)
	1.06	0.94
"I am angry"	(1.05)	(0.96)
		0.33
Avg Percent Returned (θ)		(0.04)
	4.35	2.74
Constant	(0.23)	(0.29)
	0.18	0.13
# of Obs	288	288
R^2	0.33	0.45

** p < .01, * p <.05, † p < .10; (White std. err. in parentheses)