Running Head: APLOGY AS COSTLY SIGNALING

Do Sincere Apologies Need to Be Costly?

Test of a Costly Signaling Model of Apology

Yohsuke Ohtsubo Esuka Watanabe

(Kobe University)

Published in *Evolution and Human Behavior* Vol. 30, No. 2, (2009).

Acknowledgement

This research was supported by the Japan Society for the Promotion of Science (Grant no. 18730398). We are grateful to Keiji Iwasaki, Kimiko Shiode, and Shin'itsu Tomohiro for their help in data collection, and to Martin Daly, Daniel Fessler, Tatsuya Kameda, Masanori Takezawa and anonymous reviewers for their helpful comments on earlier drafts. We would like to extend our thanks to our research assistants, Nami Adachi, Fumiko Masuda, Kazumi Nakae and Akihiro Shuda.

Abstract

The present study examined a costly signaling model of human apology. The model assumes that an unintentional transgressor is more motivated to restore the relationship with the victim than an intentional transgressor who depreciates the relationship. The model predicts the existence of a separating equilibrium, in which only sincere apologizers will pay a certain cost to restore the relationship, while dishonest apologizers will not. Accordingly, we hypothesized that the receivers of an apology would be sensitive to the cost involved in the apology. Experiments 1 and 2 were vignette experiments, in which participants imagined that they were victims of an interpersonal transgression and received either a costly or no cost apology. The costliness of the apology was manipulated by the presence of an apology gift in Experiment 1, and by inconvenience voluntarily experienced by the transgressor to make an apology in Experiment 2. In both experiments, participants found the costly apologizer to be more sincere than the no cost apologizer. Experiment 3 employed a modified dictator game, in which a fictitious partner behaved in an unfair manner and apologized to the participants. The apology cost was manipulated as a fee for sending the apology message. The results of Experiments 1 and 2 were replicated. In addition, when given a chance to send a complaint message to the unfair person, participants in the costly apology condition abstained from doing so. Implications of the study are discussed in relation to applications of the costly signaling theory to interpersonal behavior.

Keywords: Apology, Costly Signaling Theory, Handicap Principle, Sincerity

Do Sincere Apologies Need to Be Costly?

Test of a Costly Signaling Model of Apology

1. Introduction

The communicative abilities of animals/humans and their evolutionary origins have engaged the interests of many scholars from divergent perspectives (Hauser, 1997). In the animal signal literature, one of the most important issues is the reliability or honesty of signals (Zahavi & Zahavi, 1997). Honesty of signals deserves both theoretical considerations and empirical investigations, as honest communication systems are vulnerable to deceptive signalers, and thus unlikely to exist without some mechanisms to keep them honest. Zahavi's (1975) handicap principle (also known as the *costly signaling theory*) explains a mechanism whereby honesty of a signaling system becomes evolutionarily stable: High quality individuals can credibly communicate their quality by voluntarily accepting some handicaps (or cost) that low quality individuals cannot bear (see also Grafen, 1990).

Costly signaling theory has been successfully applied to some aspects of human behavior, such as altruistic behavior of hunters (e.g., Gurven, Allen-Arave, Hill, & Hurtado, 2000; Smith & Bliege Bird, 2000; Sosis, 2000), religious behavior (e.g., Irons, 2001; Sosis, 2003; Sosis & Alcorta, 2003), and human courtship behavior (e.g., Griskevicius, Tybur, Sundie, Cialdini, Miller, & Kenrick, 2007; Miller, 2000). Although some authors have suggested that costly signaling theory is also applicable to everyday interpersonal communication (Andrews, 2001; Gangestad & Thornhill, 2007), social psychological studies have paid little attention to this theory. Nonetheless, social psychologists have been interested in, and in fact have investigated, deceptive behavior in everyday interpersonal communication (e.g., DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996; Ekman, 1985). Having participants keep a diary of their deceptions, for example, DePaulo and her colleagues revealed that people tend not to tell exploitative lies, while they do often tell white lies and prosocial lies (e.g., lies told in order not to embarrass someone).

DePaulo et al.'s finding is somewhat puzzling from the perspective of the evolution of communication because human language does not involve any cost that prevents exploitative deceptions—viz., telling a lie is no more costly than telling the truth (Lachmann, Számadó, & Bergstrom, 2001; Zahavi & Zahavi, 1997). It is known that low cost signals (or cheap talk) can be honest if signalers and receivers share their interests to a substantial degree or if the signaling game has the coordination game-like incentive structure (Crawford & Sobel, 1982; Farrell & Rabin, 1996; Maynard Smith & Harper, 2003). This explanation might hold in interpersonal communication between relatives or close friends. However, people do not always share their interests. Being deceived and exploited in social exchange is considered to be a serious adaptive problem (Cosmides, 1989). Hence, it is naturally predicted that verbal communication needs to be accompanied by some costly signal when a deceptive incentive is large enough (i.e., when the honesty of low cost signals is not warranted by shared interests; Zahavi & Zahavi, 1997). As a test case, in the present study, we shall apply the costly signaling theory to human apology.

1.1. Costly Signaling Model of Apology

The apology-making context can be considered one of the situations where the reliability of signals becomes a crucial concern. If a victim unwittingly forgives a transgressor varnishing over his or her exploitative intent with verbal apology (e.g., saying "I am sorry"), he or she may be subject to similar transgressions again in the future. In this section, we shall first develop a formal model of costly apology. In the following sections, we shall briefly review previous studies related to the idea of costly apology and provide an overview of the present study.

As in the standard model of the signaling game, we assume two players: a message sender (S) and a message receiver (R). There is asymmetric information between S and R: S has private information to which R cannot have direct access. In the apology-making context, given some transgression committed by the apologizer, the private information is whether the transgression was committed accidentally. To make the model more concrete, suppose that cooperative S, denoted as S_C, has accidentally committed a personal transgression against R, and obtained some benefit, b_e from it. S_C sincerely feels sorry for it and says "I am sorry" to R for her wrongdoing (henceforth, we shall use feminine pronouns for S and masculine pronouns for R). Alternatively, the private information can be defined as whether S sincerely repented her capriciously committed transgression. In either case, it is expected that S's sincerity is correlated with the likelihood of her future cooperation.

Receiving S's apologetic statement, such as "I am sorry," R needs to be cautious because he expects to receive a similar apology not only from S_C but also from an exploitative S, denoted as S_E, who does not sincerely feel sorry. S_E might also say "I am sorry," expecting that she will be forgiven and can exploit R again. Although S personally knows which type (S_C or S_E) she actually is, R does not. In some instances, R may infer from circumstantial evidence that S committed the transgression accidentally (Malle & Knobe, 1997), and assumes that S is S_C. In other instances, R may suspect that S did it with exploitative intent, and erroneously assume S, who is in fact S_C, being S_E. In the latter instances, because merely saying "I am sorry" will not work, S_C and R need some costly signaling system that prevents S_E from producing the deceptive signal.

Assume that both S_C and R will gain the benefit of b_c from one round of cooperative interaction, and S will gain the benefit of b_e ($> b_c$) from one round of exploitation regardless of whether it was intentionally committed or not. Their interaction will be repeated with the probability of w if R decides to continue the relationship with his current partner. However, he will terminate the relationship when he thinks that the likelihood of his current partner being S_E is too high to justify the continuation of the relationship. Not to lose the potentially beneficial relationship with R, S_C has to somehow prove her true identity to R. If she successfully convinces him that she is S_C, her net benefit from interactions with R is $b_e + b_c \times w/(1 - w)$. Here, $w/(1 - w)$ is the expected number of future interactions.

Suppose that S_C pays the cost of a ($\geq b_e$) in making her apology. By definition, S_E is not willing to pay any cost greater than the benefit from the one-shot exploitation, b_e. On the other hand, S_C has an incentive to pay it if she is better off by paying the cost a to assure the future benefit of $b_c w/(1 - w)$ than keeping the benefit from one-shot exploitation, that is, $b_e \leq b_e - a + b_c w/(1 - w)$. The model can be summarized in the following inequality:

$$b_e \leq a \leq b_c w/(1 - w).$$

When the above inequality holds, S_C will make the costly apology while S_E will not. Accordingly, R can be assured that anyone paying the apology cost of a is not S_E.

The present model assumes that S_C's paying a will be offset by the benefit from repeat interactions. A similar idea was proposed by an economist, Nelson (1974), to explain the utility of dissipative advertisements, whose cost, he supposed, is offset by repeat purchases only when the advertiser produces a high quality good (see also Gintis, 2000

Chapter 13). More relevant to the apology-making context, McElreath and Boyd (2007) proposed a similar explanation for why the contrite tit-for-tat (CTFT) strategy works in the noisy repeated prisoner's dilemma game, in which players sometimes defect by mistake. CTFT accepts the partner's defection once when it has accidentally committed a defection in the previous round. McElreath and Boyd argue that a person who accidentally committed a defection can credibly signal his or her good intent by accepting the partner's defection without retaliating.

Regardless of the similarity to the above models, the present model differs from them in its assumption regarding the signaler's type. It is common in the signaling literature that S's type is modeled as her stable trait (e.g., whether one produces a high quality good is a stable trait of the producer). On the other hand, we do not necessarily assume that S's type is her trait, stable across situations and partners. If it is her trait, R may know a particular S's type not only from her signal but also from other sources, such as her reputation. We maintain that the costly apology is more important when S's type is relation-specific. Some S may find the relation with the particular R less valuable (e.g., cooperative interaction with him may bring her a benefit much smaller than b_c). Therefore, the sincerity in the present model can be defined as a proximate cue of S's valuation of the current partner, R (see Camerer's, 1988, signaling model of gift-giving for a similar assumption regarding different types of signalers; see also Appendix for preliminary evidence for this assumption). In summary, the present model assumes that the more valuable S finds a particular relationship, the more likely she is to pay a substantial amount of apology cost. She will be likely to cooperate within the relationship because she highly values it.

1.2. Other Signaling Models of Peacemaking

It is useful to compare the present model to related models in order to clarify implicit assumptions and boundary conditions of the present model. In fact, some authors argue that no cost signals work well in certain peacemaking settings. In the present section, we shall explain how the present model differs from the related models.

Silk, Kaldor and Boyd's (2000) developed a low-cost signaling model of benign disposition, and applied it to explain how female monkeys avoid hostile interactions by no cost signals (grunts and girneys). Silk et al.'s model assumes that a signaler is peaceful for p of the time, while it is hostile for $(1 - p)$ of the time. Silk et al. proved that if the receiver employs a conditional strategy (i.e., believe the signal until being deceived), the signaler's strategy of honestly signaling its current disposition is evolutionarily stable. Given the receiver's conditional strategy, a single instance of dishonest signal (i.e., disguising a peaceful disposition) will entirely eliminate the signaler's chance of future peaceful interactions with the receiver. Accordingly, insomuch as the expectation of future interactions outweighs the benefit from one-shot exploitation, every female monkey has an incentive to reveal her current disposition honestly, and thus the no cost signals can be honest. This conclusion partly depends on Silk et al.'s assumption that each S's type fluctuates over time even within a specific relationship. On the other hand, the present model assumes that it is constant over time within a specific relationship. Therefore, by definition, S_E's expected benefit from future interactions would never exceed the benefit from one-shot exploitation. As such, S_E has an incentive for dishonesty, and S_C must pay the cost, *a,* to distinguish herself form S_E. Another important difference between the present model and Silk et al.'s model is the timing of the signal (i.e., before or after a transgression). The transgression is not yet committed in Silk et al.'s model but has been committed in the present model. The timing is important because the pre-transgression signals leave more room for coordination, whereby low-cost signaling systems can be evolutionarily stable.

Recently, an economist, B. Ho (unpublished data) developed a signaling model of apology whereby he showed that no cost apology can be honest. There are at least two conceivable reasons why Ho's model and the present model arrived at different conclusions in terms of the effectiveness of no cost apologies. First, Ho's model deals with R's behavioral reaction toward various apologies, while the present model focuses on R's perception of sincerity. The present model can be considered as a subset of Ho's model (i.e., R's Bayesian inference process upon receiving an apology). According to the Bayesian game framework (Harsanyi, 1967-1968), R is supposed to possess some prior probability of S being S_C. Observing S's transgression (i.e., S_E-like behavior), R is likely to lower his subjective probability of S being S_C. S's apology is thus considered to influence R's subjective probability. According to this framework, whether R forgives S is jointly determined by (1) R's subjective probability of S being S_C, (2) his potential benefit from his relationship with S, and (3) his potential loss from a further exploitation by S. Our model exclusively focuses on the first factor (i.e., R's subjective probability) referring to it as sincerity perception. On the other hand, if the focus is whether R will forgive S, it is logically predicted that cheap apologies are sufficient when the potential benefit is large and the potential loss is negligible (see Appendix for preliminary evidence that R's perception of sincerity and willingness to forgive are dissociable under a certain condition).

The second difference between Ho's model and the present model is the assumption regarding the presence of some exogenous cost. Even when R is suspicious about S's type, the no cost apology equilibrium can be sustained by the presence of exogenous cost, such as punishment against deceptive apologies (B. Ho, unpublished data). We agree with Ho that the presence of exogenous cost plays an important role to sustain honesty in human communication (Lachmann et al., 2001; Y. Ohtsubo, F. Masuda, & E. Watanabe, unpublished data). However, we are dubious about the effectiveness of punishment in the apology-making context. Suppose that a receiver accepts a no cost apology but threatens the signaler by stating that "I will punish you if you do it again (i.e., if the apology turns out to be dishonest)." The signaler may not take the receiver's threat seriously because the receiver signals his disinclination to punishment by failing to inflict it exactly when the threat is being conveyed (Schelling, 1960).

1.3. Overview of the Present Study

The costly signaling model of apology predicts that people in the victim's perspective are sensitive to the presence of the apology cost in determining the apologizer's sincerity. Accordingly, it predicts that people will be more likely to forgive a costly apologizer than a no cost apologizer, although we have noted that whether one perceives the apologizer as sincere cannot be equated with whether one forgives the apologizer. Supportive evidence has been accumulated by social psychological research on apology: if a transgressor voluntarily offers some compensation or gift, he or she is more likely to be forgiven (Darby & Schlenker, 1982; Eaton, Struthers, & Santelli, 2006; Gauché & Mullet, 2005). A similar observation was made by a primatologist, de Waal (1989), who noted that gift-giving is a uniquely human way of peacemaking. Notice that an apologizer inevitably pays some cost in making compensation or giving a gift to a victim. We first tested the effect of the apology gift on the *sincerity perception* in a vignette experiment (Experiment 1).

The effect of the apology gift (or compensation) is, however, subject to an alternative explanation. Because the cost paid in the form of a gift or compensation is transferred to the victim, the materialistic value of a costly apology is greater for the victim than that of a no cost apology. Accordingly, those who received a costly apology might perceive the apology more positively (i.e., more sincere) because they are happier than those who received a no cost apology. This problem pertains not only to Experiment 1 but also to the previous research revealing positive effects of apology gifts and compensation. Our model, nonetheless, predicts that the cost *per se* is crucial for a victim to evaluate the apologizer's sincerity. Another vignette experiment, Experiment 2, was thus conducted to test the prediction that apology cost that is not transferred to a victim is sufficient for him or her to perceive the sincerity in the apology. For this purpose, we devised scenarios describing the cost such that it would reduce the transgressor's payoff but not increase the victim's payoff (e.g., apologizer's canceling a part-time job to make an apology as soon as possible).

Experiment 3 tested the same prediction with a behavioral experiment. In Experiment 3, participants played a modified version of the dictator game with a fictitious partner (see Camerer & Fehr, 2004, for an explanation of the standard dictator game). Participants learned that the partner allocated them only 200 of 1000 JPY, and then apologized for the unfair allocation. Costliness of the apology was experimentally manipulated as a fee to send an apology message. The cost paid by the dictator was not transferred to participants. In addition to the perceived sincerity, Experiment 3 assessed participants' behavioral reaction to the costly/no cost apologizer by asking them to indicate their willingness to send a complaint message to the partner. Recall that the present model does assume that the perceived sincerity influences R’s behavioral reaction, while it does not assume that the perceived sincerity is the sole determinant of the behavioral reaction.

One might criticize Experiment 3, noticing that it violated an assumption of the model—viz. participants did not engage in the repeated interactions. We consider that the exclusion of the repeat interactions was a rather desirable feature to test the present, evolutionary hypothesis. There are three recent studies that compared the effect of costly/no cost apology in the repeated economic game settings (Bottom, Gibson, Daniels, & Murnighan, 2002; B. Ho, unpublished data; Schweitzer, Hershey, & Bradlow, 2006). Interestingly, they differ in their conclusions regarding the effect of no cost apology. Investigating the effect of different types of apology on trust restoration, Schweitzer et al. found little facilitative effect of the no cost apology. Bottom et al. found a small but positive effect of no cost apology. More importantly, however, Bottom et al.’s study revealed that costly apology was more effective than no cost apology. Ho also found a positive effect of no cost apology, and did not find a significant effect of the level of apology cost. The contradictory results might be partly attributable to the methodological features of the previous studies. All three studies measured the apology receivers’ behavioral reaction in some repeated game setting. The repeated game allowed the participants to behave according to their prospects of success in the game, and thus whether no cost apologies were effective might have depended on the specific game structure that those researchers employed in their experiments. This argument suggests that forgiving S who made a no cost apology might have been a product of rational calculation in these studies. Exclusion of the repeat interactions from Experiment 3 allowed us to eliminate variants of rational choice explanations for the observed results.

Moreover, we argue that the present model in fact predicts the effect of apology cost on the perceived sincerity in the absence of the repeat interactions. The present model's specific prediction is that people have some innate sensitivity to the cost involved in apologetic signals. Such a cost sensitive psychology is supposed to have evolved in the environment of evolutionary adaptedness (EEA), in which the repeat interactions along with the other game structures existed. Once evolved in the EEA, our innate psychological mechanisms are triggered by proximate cues that were correlated with the hypothesized game structures in the EEA (Hagen & Hammerstein, 2006; Haley & Fessler, 2005). We consider that plausible proximate cues for the sincerity assessment are the contextual frame (i.e., apology-making context) and cost involved in apologetic signals. We do not consider the repeat interactions as a plausible proximate cue for the sincerity assessment. In the EEA, most social interactions probably occurred among the same community members, thus most interactions were repeated (see Maynard Smith & Harper, 2003, p. 126, for a similar argument). Something that is always present cannot serve as a good criterion on which one bases his or her decision. Accordingly, we considered that Experiment 3 was rather a suitable setting to test the evolutionary hypothesis derived from the costly signaling model.

2. Experiment 1

2.1. Method

2.1.1. Participants and Design. Participants were 60 undergraduates (23 males and 37 females) at a small private university in Japan. They volunteered to participate in a set of experiments for a monetary reward (1000 JPY). The costliness of apology (costly apology vs. no cost apology) was a between-participants factor: 30 participants were assigned to the costly apology condition, and the other 30 to the no cost apology condition. This study was followed by an unrelated experiment. Each experimental session involved a maximum of eight participants.

2.1.2. Materials. Participants were presented with a fictitious scenario describing a situation in which their friend borrowed an electronic dictionary from them and failed to return it. The scenario read that they suffered inconvenience from the unavailability of the dictionary in completing their own homework. After a week, the friend returned the dictionary to them. There were two versions of the apology scenario. In the no cost apology condition, the friend apologized to them, saying "I'm sorry that I unwittingly forgot to return it to you." In the costly apology condition, in addition to the apology statement, the friend bought them lunch. The buying lunch was employed as the manipulation of apology cost because it was most frequently mentioned in a pilot study that asked a different sample of participants drawn from the same population, by means of an open-ended questionnaire, how they would apologize if they were in the transgressor's position.

2.1.3. Dependent Variables. Participants first read the transgression scenario, and rated their anger on a 5-point scale (1: not at all – 5: very much). This item was included to confirm that there were no systematic differences between the conditions prior to the apology manipulation. Participants then read one of the apology scenarios according to their condition. After reading it, they answered a questionnaire, in which the perceived sincerity item was embedded. Participants rated how sincere they found the transgressor's apology (i.e., *perceived sincerity*) on a 5-point scale. Although there were several other related items, we shall only report the analyses of the perceived sincerity because it is most relevant to the model's prediction.

2.2. Results

The mean±s.d. pre-apology anger score was 3.87±0.94 (*n* = 30) in the no cost apology condition and 4.07±0.69 (*n* = 30) in the costly apology condition, t_{58} = .94, *ns*. Therefore, before the apology manipulation, the anger score was comparable across the conditions. As predicted, the mean perceived sincerity was significantly higher in the costly apology condition, t_{58} =3.43, *p* = .001 (one-tailed test), *d* = 0.88 (see Figure 1).

3. Experiment 2

3.1. Method

3.1.1. Participants and Design. Participants were 77 undergraduates (46 males and 31 females) at a small private university in Japan. The study was conducted in a large classroom as a part of the course requirements. The costliness condition was a between-participants factor. To test the generalizability of the results, three new vignettes were written and included as repeated measures. All three scenarios presented to each participant were from the same condition, either the no cost apology or costly apology condition. The scenario order was kept constant for all the participants. Experiment 2 also employed the perceived sincerity as the dependent variable.

3.1.2. Vignettes and Manipulations. In this section, we shall explain the nature of the three vignettes using one scenario. (English versions of the three vignettes are available from the correspondence author upon request.) The vignette described a situation in which the participant's same-sex friend changed the shift of a part-time job without consulting the participant because he or she had to go to his or her hometown suddenly. Accordingly, the participant was scheduled to work on a Friday. The participant would not normally have refused to work on Friday. However, the participant had an important test on the weekend. Because the participant worked on the Friday night, he or she was not able to finish preparation for the test. The apology scenario contained an account stating that the friend inadvertently forgot about the test. The other vignettes shared the following features: every vignette described a same-sex friend who gained (or tried to gain) some benefit (e.g., substitution for the part-time job) from the transgression; in every vignette, the participant was unable to forestall the friend's action.

The costliness of the apology was manipulated in the following manner: In the costly apology condition, the scenario read that after noticing his or her transgression, the friend changed his or her plans (i.e., canceled a flight originally reserved and bought a new ticket) and hurriedly came to the participant to make an apology. In the no cost apology condition, the friend came back according to his or her original schedule, and apologized to the participant at his or her first incidental encounter with the participant after the transgression. For the other two vignettes, the costliness was manipulated as follows: for one vignette, the friend canceled his or her part-time job to make an apology as soon as possible; for the other vignette, the friend showed up to an early morning class, in which he or she was not enrolled, just to make an apology.

3.2. Results

We report the results based on the average of the three vignettes. We shall report only statistics associated with the aggregate results to save space. The results of the separate analyses for the three vignettes were all comparable to those of the aggregated analyses.

The mean±s.d. pre-apology anger score was 2.88±0.78 (n = 40) in the no cost apology condition and 3.07±0.83 (n = 37) in the costly apology condition, respectively, t_{75} = 1.05, *ns*. Therefore, it can be considered that the two conditions were comparable before the apology manipulation. One might notice that the pre-apology anger score was considerably lower in Experiment 2 than in Experiment 1. This may be partly attributable to the fact that all scenarios in Experiment 2 described indirect transgressions so that the fictitious costly apologies (i.e., delayed apologies) would make sense to participants.

Cronbach's coefficient α associated with the perceived sincerity was .76 for the three vignettes. As was observed for Experiment 1, the perceived sincerity was higher in the costly apology condition than in the no cost apology condition, $t_{75} = 4.56$, $p < .001$ (one tailed test), $d = 1.04$ (see Figure 1). Confirming the effect of costly apology with the two vignette experiments, we proceeded to test the effect with a behavioral measure.

4. Experiment 3

4.1. Method

4.1.1. Participants. Participants were 42 (19 males and 23 females) undergraduates who were enrolled in an evolutionary psychology class at a small private university in Japan. They participated in the study for partial credit in the course and a small monetary reward (i.e., 200 JPY), whose exact amount had not been specified in advance.

4.1.2. Procedure. Forty two participants were assembled in a large room. Each participant received an ID card on which his or her ID number was printed either in red or blue ink. Participants were divided into two groups of 21 people and re-assembled in two rooms according to the color of their ID numbers. In both rooms, participants were assigned ID numbers 1 through 21, and instructed that they were paired with another participant assigned to the same ID number, printed in a different color. It was then explained to the participants that they would play the dictator game in the receiver role, while their partner would play the game in the dictator role. However, all participants in fact played the receiver role. Throughout the instructions to participants, more neutral words, "allocation game" and "allocator," were used instead of "dictator game" and "dictator."

The dictator game, which was modified for the purpose of the present study, was explained to participants as follows: The dictator would be asked to allocate 1000 JPY between the receiver and him- or herself. There were nine possible choices of allocation, each giving the participant 100 through 900 JPY with increments of 100. However, the dictator would be allowed to choose from only two allocation schemes that had been randomly predetermined by a computer program for each ID number. The experimenter clearly stated that the dictator might give the receiver more or less than 500 JPY because of the limited choice (e.g., they might be assigned two allocation schemes, each giving 300 and 400 JPY to the receiver).

After the instruction, the experimental assistant brought envelopes to each room, apparently from the other room. The experimenter distributed the envelopes according to the participants' ID numbers. All envelopes contained a slip showing that the dictator kept 800 JPY for him- or herself and gave only 200 JPY to participants. After inspecting it, participants answered Questionnaire 1, in which a pre-apology anger item was embedded among filler items.

When all participants had completed Questionnaire 1, the experimenter explained to the participants that the dictator was allowed to apologize to them if he or she had allocated less than 500 JPY to the participants. The experimenter explained that there were two ways to make an apology: (i) the dictator could write his or her apology message (within a limitation of 25 Japanese letters) if he or she agreed to pay 500 JPY, which was collected by the experimenter and not transferred to the participants; (ii) alternatively, the dictator could choose an apology message from several ready-made messages and send it for free. It was explained to the participants that there were several apology messages frequently written by former participants in similar experiments. Accordingly, the experimenter prepared several stereotypical apology messages for the dictator. The contents of ready-made messages were not specified to the participants. After the explanation of the apology message, the experimenter distributed envelopes to the participants according to their ID number. To conceal the manipulation of the study, the experimenter announced that the envelope would contain a blank sheet for those who had received 500 JPY or more. In reality, approximately half of the participants in each room received the costly apology, while the other half received the no cost apology. In both conditions, the apology message read: “I am sorry but I had no choice.” It was explained that the dictator who chose to send a ready-made message had to copy it onto the apology sheet by him- or herself. Accordingly, participants in both conditions received an apparently identical hand-written apology message. The apology sheet had two distinct spaces, one for the no cost apology and the other for the costly apology. The space for the costly apology was marked by the following statement, which was printed in red ink ostensibly to give the dictator a caution: “You have to pay 500 JPY to write an apology message by yourself.” Because one of the two spaces was left blank, participants in both conditions readily recognized which apology message they received. After receiving either a costly or no cost apology message, participants answered Questionnaire 2 containing the perceived sincerity item, which were embedded among filler items. It was clearly stated on the questionnaire that the sincerity item was intended to be answered by only those who received an apology message.

After completing Questionnaire 2, participants were further instructed that they could send a complaint message. Participants were given a complaint sheet on which they were asked to indicate whether they would like to send a complaint to the dictator. If they decided to send it, they were further asked to choose one complaint message from four options. The four messages were as follows: (i) “I am angry with you,” (ii) “You ought to feel guilty,” (iii) “I don’t know why you did it,” and (iv) “I wanted you to behave more fairly.” These four options were included to make the task realistic. This task in fact measured participants’ willingness to express complaint as a binary variable. We predicted that if participants considered they had received a sincere apology, they would no longer want to express complaint. After all participants completed the complaint sheet, the experimenter briefly explained the nature of the experiment and paid 200 JPY to each participant. More thorough feedback was provided by the first author later in class.

Experiment 3 included deceptions. We admit that deceptions could have some detrimental effects on participants’ trust in psychological/economic experiments (Ortmann & Hertwig, 2002). However, we considered that, for the present purpose, some deceptions were necessary. For some constraints, we had to recruit participants from a single class, in which many mutual friends were enrolled. If participants had in fact engaged in some unfair allocation and apologized to an anonymous partner, the partner might happen to know the identity of the transgressor from his or her handwriting. Such an incidence could have a detrimental effect on participants’ real friendships. To avoid this potential problem, we decided to employ deceptions.

4.2. Results

The mean±s.d. pre-apology anger scores were 2.71±1.42 ($n = 21$) in the no cost apology condition and 2.71±1.10 ($n = 21$) in the costly apology condition, respectively, $t_{40} = 0$, *ns*. Therefore, there was no difference across conditions before the apology manipulation. The mean perceived sincerity was significantly higher in the costly apology condition, $t_{40} = 3.58$, $p < .001$ (one-tailed test), $d = 1.11$ (see Figure 1). This result replicated those of the two vignette experiments. Experiment 3 included another dependent variable: Whether the participant sent the complaint message to the dictator. In the no cost apology condition 7 of 21 participants sent it, while in the costly apology condition only 1 of 21 participants did so, $p = .022$ by Fisher's exact test (one-tailed). It is noteworthy that the participant who sent the massage in the costly apology condition spontaneously left a statement to the effect that "you won't pay the cost anyway" on the complaint sheet. Therefore, if we consider that the manipulation was unsuccessful for this participant, the complaint message was not sent by any participants who believed that they received a truly costly apology.

5. General Discussion

The present study demonstrated that a costly apology effectively communicates the apologizer's sincerity to the victim. In Experiment 1, we showed that the apology with a gift was more effective in communicating the apologizer's sincerity than the apology without it. In Experiment 2, we showed that an apology cost that was experienced as inconvenience merely to the apologizer was also effective in communicating sincerity. Experiment 3 tested the effect of a costly apology in an experimental setting in which a fictitious partner made the participant angry by allocating a monetary reward in an unfair manner and then apologized, stating that his or her unfair allocation had not been intended. Experiment 3 confirmed the same pattern as the two vignette experiments: Participants found the apology sincerer in the costly apology condition. Furthermore, Experiment 3 revealed that participants who received a costly apology were less likely to express their complaint to the fictitious partner than those who received a no cost apology. Of particular importance is that Experiments 2 and 3 verified the effect of cost *per se*, rather than the effect of materialistic benefits that the apologized party would receive in the case of an apology gift or compensation.

Admittedly, the findings of this series of experiments do not eliminate alternative explanations. Thus, we shall briefly consider two of the alternative hypotheses and present some preliminary data against them. First, a victim may regard costly apologizers as good-natured people who are unlikely to engage in a transgression intentionally. We shall refer to this hypothesis as the *good-person hypothesis*. The second explanation is based on an assumption that people are concerned not only with their own absolute well-being, but also with the relative well-being between themselves and others (Crosby, 1976; Fehr & Schmidt, 1999). Observing an apologizer pays some cost, the victim might perceive a restored balance between him or her and the apologizer. Accordingly, participants might have perceived the costly apology more positively because they were in a happier mood by seeing the restored balance. We shall refer to this hypothesis as the *restored balance hypothesis*.

To eliminate the good-person hypothesis, we conducted a small-size follow-up experiment similar to Experiment 3. In the new experiment, participants were led to believe that the dictator was playing the game with six partners. Each participant received feedback indicating that there were two participants, including him- or herself, who received only 200 of 1000 JPY. In the costly apology condition, the participant was told that the dictator made a costly apology to him or her but made a no cost apology to the other participant who had also received only 200 JPY. In the no cost apology condition, the participant was told that the dictator made a no cost apology to him or her but made a costly apology to the other participant. Therefore, objectively, the dictator's goodness was constant between the two conditions. The mean perceived sincerity score was still higher in the costly apology condition, 4.56±0.53, than in the no cost apology condition, 3.45±1.29, $t_{18} = 2.39$, $p = .028$, $d = 1.13$.

To eliminate the restored balance hypothesis, we conducted a vignette experiment, in which there were two additional conditions to the costly and no cost apology conditions. The first condition was a good luck condition that asked participants to imagine that they had a lucky experience because of the no cost apologizer's transgression: Right after the transgressor returned the participant's textbook, which he or she had taken without the participant's permission, the participant saw a professor. The professor happened to notice that the participant held the textbook and praised him or her for his or her diligence. This event was considered to be relevant to the restored balance hypothesis because the scenario read that the professor had rebuked the participants who had been present in the class without the textbook. The second condition was a bad-luck condition that asked participants to imagine that the transgressor was stranded at the station for an hour due to a train accident. This condition was relevant to the restored balance hypothesis because the one-hour wait matched the apology cost in the costly apology condition. The costly apology scenario read that in order to make an apology, the transgressor had been waiting for the participant for an hour outside of the participant's workplace while the participant was working for some extra time beyond his or her ordinary working hour. A one-way ANOVA revealed the effect of the condition on the perceived apology, $F_{3,\ 89} = 11.75$, $p < .001$, $\eta_p^2 = 0.28$. Post hoc comparisons indicated that the mean perceived sincerity score was significantly higher in the costly apology condition (4.67±0.48) than in the other three conditions (3.83±0.73, 4.00±0.80, and 3.41±0.85 for the no cost, good luck, and bad luck conditions, respectively).

Although we showed that the perceived sincerity associated with the costly apology was not fully explained by the good-person hypothesis, a recent study provided some support for it: Victims tend to infer that transgressors are more agreeable if they made amends (B. A. Tabak & M. E. McCullough, unpublished data). This is not totally surprising, given that appropriate signal costs can communicate the signaler's stable traits as well as ephemeral mental states. These two types of signaling models differ in the assumption of how the cost is offset. According to the good-person hypothesis, costly apologizers will benefit from a good reputation (e.g., they might enjoy cooperative interactions with a greater number of partners than no cost apologizers; cf. Frank, 1988; Roberts, 1998). In such a model, the signals are targeted not only to a current interaction partner but also to a wider range of audience. The present model assumes that the apology message was exclusively targeted to the receiver. The honest signaler, S_C, will benefit from retaining a cooperative relationship with R. Admittedly, apology costs could serve for both purposes among humans. To fully comprehend human apology as a costly signal, we have to study not only the victim's perception of the costly/no cost apologies but also the third parties' perception of them. Also, we need to extend our focus to include the apologizers' perspective, and investigate what conditions facilitate/prevent apologizers' paying the apology cost.

Some evidence suggests that people who feel guilty are willing to pay a cost to benefit the victim (see Schlenker, & Darby, 1981; D. Sznycer, J. G. Price, J. Tooby, & L.

Cosmides, unpublished data). It is thus possible that what we have called sincerity in the present study is more appropriately understood as the sense of guilt felt by the transgressor. Consistent with the present model's assumption, Baumeister, Stillwell, and Heatherton (1994) noted that people tend to experience a stronger sense of guilt within close relationships (perhaps more valuable relationships), although they also noted that some people are more prone to feel guilty than others (i.e., support for the good-person hypothesis). Further studies are needed to clarify the relation between guilt and sincerity in the apology-making context. Because both guilt and sincerity are mental states of the signaler, this line of research will illuminate the importance of the signaling game framework in understanding the phenomena embraced under the rubric of mind-reading. The traditional research on mind-reading tended to conceptualize it as each individual's cognitive processes (e.g., Ames, 2004). On the other hand, costly signaling theory conceptualizes it as collaboration between players (i.e., signalers and receivers). Therefore, costly signaling theory forces us to study the behavioral strategies of the being read (i.e., signaler), let alone those of the mind reader.

The present model's framework may also contribute to an evolutionary understanding of mind-reading and trait inference abilities. Mind-reading ability has often been modeled as a counterpart of manipulation (e.g., Krebs & Dawkins, 1984). This conceptualization of mind-reading implicitly assumes that the relation between signalers and receivers is something like the zero-sum game, in which one party's gain necessarily implies the other party's loss. On the other hand, the present model assumes that signalers and receivers share some of their interests in common (cf. Schelling, 1960; Zahavi & Zahavi, 1997). Because of the difference in the assumed game structure (i.e., zero-sum game vs. mixed motive game), the two ways of conceptualization seem to be associated with different predictions concerning evolved abilities of mind-reading. Empirical tests of such predictions may facilitate our understanding of the evolution of mind-reading. In summary, it seems fertile to apply the costly signaling theory to study interpersonal behavior and underlying social cognition.

Appendix

We have collected some preliminary data showing that R's willingness to forgive can be dissociated from the perceived sincerity. We conducted a follow-up study that was same as Experiment 2 except the following two modifications: First, we included three additional dependent variables (i.e., willingness to forgive, inference of the friend's valuation of the relationship, inference of the friend's sense of guilt). Second, each scenario included an additional statement that the friend did not have adequate knowledge to foresee that his or her action would have some harmful effect on the participant (e.g., the friend did not know that the participant was planning to take a test). The lack of relevant knowledge is a sufficient condition for people to infer that the harmful effect had not been intended (Malle & Knobe, 1997). Despite this modification, the mean perceived sincerity was higher in the costly apology condition (4.20±0.78) than in the no cost apology condition (3.70±0.66), $t_{68} = 2.90$, $p = .003$, $d = 0.69$. The two additional variables, the inference of friend's valuation and sense of guilt, also showed a consistent pattern: The mean ratings were higher in the costly apology condition than in the no cost apology condition, 4.12±0.67 vs. 3.52±0.76, $t_{68} = 3.50$, $p = .001$, $d = .84$ for the valuation, and 4.15±0.70 vs. 3.75±0.68, $t_{68} = 2.39$, $p = .020$, $d = 0.57$ for the guilt. On the other hand, the willingness to forgive score did not significantly differ between the conditions, 4.06±0.78 vs. 3.83±0.81 for the costly apology and no cost apology conditions, respectively, $t_{68} = 1.23$, *ns*.

References

Ames, D. R. (2004). Inside the mind reader's tool kit: Projection and stereotyping in mental state inference. *Journal of Personality and Social Psychology*, **87**, 340-353. (DOI 10.1037/0022-3514.87.3.340)

Andrews, P. W. (2001). The psychology of social chess and the evolution of attribution mechanisms: Explaining the fundamental attribution error. *Evolution and Human Behavior*, 22, 11-29. (DOI 10.1016/S1090-5138(00)00059-3)

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, **115**, 243-267. (DOI 10.1037/0033-2909.115.2.243)

Bottom, W. P., Gibson, K., Daniles, S. E., & Murnighan, J. K. (2002). When talk is not cheap: Sbustantive penance and expressions of intent in rebuilding cooperation. *Organization Science*, **13**, 497-513. (DOI 10.1287/orsc.13.5.497.7816)

Camerer, C. (1988). Gifts as economic signals and social symbols. *American Journal of Sociology*, **94**, S180-S214.

Camerer, C. F., & Fehr, E. (2004). Measuring social norms and preferences using experimental games: A guide for social scientists. In J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Ferh, & H. Gintis (Eds.), *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies* (pp. 55-95). Oxford, England: Oxford University Press.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, 31, 187-276. (DOI 10.1016/0010-0277(89)90023-1)

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, **50**, 1431-1451.

Crosby, F. (1976). A model of egoistical relative deprivation. *Psychological Review*, **83**, 85-113. (DOI 10.1037/0033-295X.83.2.85)

Darby, B. W., & Schlenker, B. R. (1982). Children's reactions to apologies. *Journal of Personality and Social Psychology*, 43, 742-753. (DOI 10.1037/0022-3514.43.4.742)

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, **70**, 979-995. (DOI 10.1037/0022-3514.70.5.979)

Eaton, J., Struthers, C. W., & Santelli, A. G. (2006). The mediating role of perceptual validation in the repentance–forgiveness process. *Personality and Social Psychology Bulletin*, 32, 1389-1401. (DOI 10.1177/0146167206291005)

Ekman, P. (1985). *Telling lies: Clues to deceit in the marketplace, politics, and marriage*. New York: Norton.

Farrell, J., & Rabin, M. (1996). Cheap talk. *Journal of Economic Perspectives*, **10**, 103-118.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, **114**, 817-868.

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions*. New York: Norton.

Gangestad, S. W. & Thornhill, R. (2007). The evolution of social inference processes: The importance of signaling theory. In J. P. Forgas, M. G. Haselton, & W. von Hippell (Eds.), *Evolution and the Social Mind*, (pp. 33-48). New York: Psychology Press.

Gauché, M., & Mullet, E. (2005). Do we forgive physical aggression in the same way as that we forgive psychological aggression? *Aggressive Behavior*, 31, 559-570. (DOI

10.1002/ab.20108)

Gintis, H. (2000). *Game theory evolving: A problem-centered introduction to modeling strategic interaction*. Princeton, NJ: Princeton University Press.

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, **144**, 517-546. (DOI 10.1016/S0022-5193(05)80088-8)

Griskevicius, V., Tybur, J. M., Sundie, J. M., Cialdini, R. B., Miller, G. F., Kenrick, D. T. (2007). Blatant benevolence and conspicuous consumption: When romantic motives elicit strategic costly signals. *Journal of Personality and Social Psychology*, **93**, 85-102. (DOI 10.1037/0022-3514.93.1.85)

Gurven, M., Allen-Arave, W., Hill, K., & Hurtado, M. (2000). “It’s a wonderful life”: signaling generosity among the Ache of Paraguay. *Evolution and Human Behavior*, 21, 263-282. (10.1016/S1090-5138(00)00032-5)

Hagen, E. H., & Hammerstein, P. (2006). Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical Population Biology*, **69**, 339-348. (DOI 10.1016/j.tpb.2005.09.005)

Haley, K. J., & Fessler, D. M. T. (2005). Nobody’s watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, **26**, 245-256. (DOI 10.1016/j.evolhumbehav.2005.01.002)

Harsanyi, J. C. (1967-1968). Games with incomplete information played by “Bayesian” players, I-III. *Management Science*, **14**, 159-182, 320-334, 486-502. (DOI 10.1287/mnsc.1040.0270)

Hauser, M. D. (1997). *The evolution of communication*. Cambridge, MA: MIT Press.

Irons, W. (2001). Religion as a hard-to-fake sign of commitment.” In R. M. Nesse, R. M.

(Ed.), *Evolution and the capacity for commitment* (pp. 292-309). New York: Russell Sage Foundation.

Krebs, J. R., & Dawkins, R. (1984). Animal signals: Mind-reading and manipulation. In J. R. Krebs and N. B. Davies (Eds.), Behavioural ecology: An evolutionary approach (pp. 380-402). Oxford, England: Blackwell.

Lachmann, M., Számadó, S., & Bergstrom, C. T. (2001). Cost and conflict in animal signals and human language. *Proceedings of the National Academy of Sciences, USA*, **98**, 13189-13194. (DOI 10.1073/pnas.231216498)

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, **33**, 101-121. (DOI 10.1006/jesp.1996.1314)

Maynard Smith, J., & Harper, D. (2003). Animal signals. Oxford, U.K.: Oxford University Press.

McElreath, R., & Boyd, R. (2007). *Mathematical models of social evolution: A guide for the perplexed*. Chicago: University of Chicago Press.

Miller, G. (2000). *The mating mind: How sexual choice shaped the evolution of human nature*. New York: Doubleday.

Nelson, P. (1974). Advertising as information. *Journal of Political Economy*, **82**, 729-754. (DOI 10.1086/260231)

Ortmann, A., & Hertwig, R. (2002). The costs of deception: Evidence from psychology. *Experimental Economics*, **5**, 111-131. (DOI 10.1023/A:1020365204768)

Roberts, G. (1998). Competitive altruism: from reciprocity to the handicap principle. *Proceedings of the Royal Society of London B*, **265**, 427-431. (DOI 10.1098/rspb.1998.0312)

Schelling, T. C. (1960). *The strategy of conflict*. Cambridge, MA: Harvard University Press.

Schlenker, B. R., & Darby, B. W. (1981). The use of apologies in social predicaments. *Social Psychology Quarterly*, 44, 271-278. (DOI 10.2307/3033840)

Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes*, **101**, 1-19. (DOI 10.1016/j.obhdp.2006.05.005)

Silk, J. B., Kaldor, E., & Boyd, R. (2000). Cheap talk when interests conflict. *Animal Behaviour*, **59**, 423-432. (DOI 10.1006/anbe.1999.1312)

Smith, E. A., & Bliege Bird, R. L. (2000). Turtle hunting and tombstone opening: Public generosity as costly signaling. *Evolution and Human Behavior*, 21, 245-261. (DOI 10.1016/S1090-5138(00)00031-3)

Sosis, R. (2000). Costly signaling and torch fishing on Ifaluk atoll. *Evolution and Human Behavior*, **21**, 223-244. (DOI 10.1016/S1090-5138(00)00030-1)

Sosis, R. (2003) Why aren't we all Hutterites? Costly signaling theory and religious behavior. *Human Nature*, **14**, 91-127.

Sosis, R. & Alcorta, C. (2003). Signaling, solidarity, and the sacred: The evolution of religious behavior. *Evolutionary Anthropology*, **12**, 264-274. (DOI 10.1002/evan.10120)

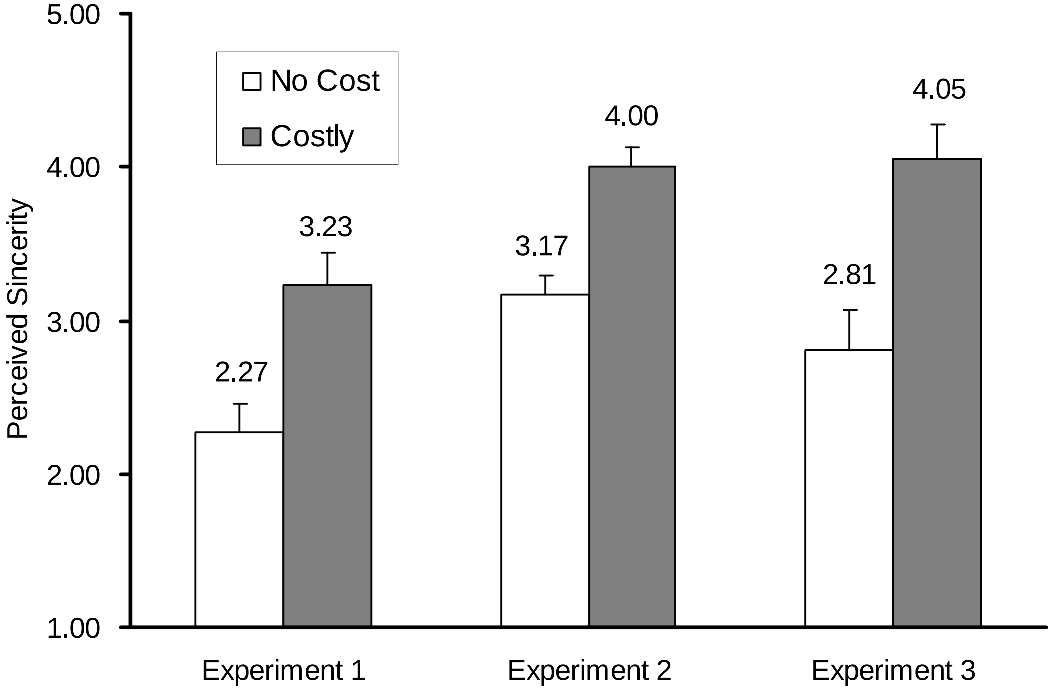
de Waal, F. B. W. (1989). *Peacemaking among primates*. Cambridge, MA: Harvard University Press.

Zahavi, A. (1975). Mate selection: A selection for a handicap. *Journal of Theoretical Biology*, **53**, 205-214. (DOI 10.1016/0022-5193(75)90111-3)

Zahavi, A., & Zahavi, A. (1997). *The handicap principle: A missing piece of Darwin's* *puzzle*. New York: Oxford University Press.

Figure Captions

Figure 1. Mean perceived sincerity as a function of the costliness of apology and experiment. Error bars indicate standard errors of means.


5.00
4.00
3.00
2.00
1.00
Perceived Sincerity
□ No Cost
■ Costly
2.27
3.23
3.17
4.00
2.81
4.05
Experiment 1
Experiment 2
Experiment 3

<Figure 1>