Costly Apology and Self-Punishment after an Unintentional Transgression

E. Watanabe & Y. Ohtsubo

(Kobe University)

Yohsuke Ohtsubo

Department of Psychology, Faculty of Letters

Kobe University

1-1, Rokkodai-cho, Nada-ku, Kobe, 657-8501, Japan

Phone: Int+81-72-803-5519

e-mail: yohtsubo@lit.kobe-u.ac.jp

Manuscript Accepted for Publication in *Journal of Evolutionary Psychology*

Abstract

Making a costly apology or inflicting self-punishment after an unintentional transgression can serve as a costly signal of the transgressor's benign intention. In the present research, after an unintentional transgression (i.e., unequal resource allocation between themselves and a partner), participants were provided with an opportunity to send an apology message to their partner (in Experiments 1 and 2) or to privately deduct some amount from their own reward (in Experiments 2 and 3). Across these experiments, approximately half of the participants indicated their willingness to incur some cost to produce these costly signals. In Experiment 1, neither history nor expectation of interaction with a partner altered the frequency of a costly apology. In Experiment 2, despite explicit instructions that their partner would not be informed whether they had inflicted the self-punishment, the frequency of self-punishment was approximately equal to that of a costly apology. These results suggest that the two types of costly signal were not solely directed at the victim. Experiment 3 revealed that these costly signallers endorsed the equality principle more than the non-signallers. This result is consistent with the idea that the two forms of costly signals serve to protect the signaller's reputation as a fair person.

Keywords: Costly Signalling Theory, Apology, Self-Punishment, Indirect Reciprocity

1. Introduction

To maintain cooperation in noisy environments, evolutionarily stable strategies need to include some mechanisms to restore mutual cooperation after a single player's mistakenly chosen defection. In the iterated prisoner's dilemma game, the contrite tit-for-tat strategy, which is similar to the standard tit-for-tat strategy except that it accepts the partner's defection after its own erroneous defection, was shown to be evolutionarily stable in the presence of implementation errors (SUGDEN 1986; BOYD 1989; WU and AXELROD 1995). The strength of the contrite tit-for-tat strategy lies in the fact that an erroneous non-cooperator's apologetic gesture (i.e., acceptance of the partner's defection) leads to the partner's forgiveness, and the pair can return to a mutually cooperative equilibrium.

A slightly different mechanism of "apology and forgiveness" is also required in the indirect reciprocity context, where people must decide whether to behave cooperatively toward a new partner based on that partner's standing (NOWAK and SIGMUND 1998). Through exhaustive analyses of 4096 possible strategies, OHTSUKI and IWASA's (2006) found that only eight strategies are evolutionarily stable in the indirect reciprocity game. These eight strategies (collectively referred to as the "leading eight") were all apologetic (i.e., all the eight strategies behave in an unconditionally cooperative manner when their standing is "bad") and forgiving (i.e., they will treat ex-bad players as "good" players if they expressed their repentance through unconditional cooperation).

The so-called "apology" in the above strategies takes the form of unconditional cooperation because such a costly action can honestly signal the apologiser's benign intention. Empirical evidence confirms the above models' assumption that people behave cooperatively after defection. In KETELAAR and AU's (2003) study, for example, participants who behaved non-cooperatively in an economic game and then felt guilty became more cooperative toward the same partner in a subsequent economic game. Such apparent reparation also extends

beyond the dyadic relation (i.e., the direct reciprocity context). MILINSKI, SEMMANN, BAKKER and KRAMBECK (2001), for example, revealed that participants who had justifiably punished a "bad" player by choosing defection compensated for their justifiable defection, which might have been misperceived as bad behaviour, by increasing cooperation toward other "good" players. A similar pattern has been observed in field experiments, in which participants who committed a transgression (e.g., breaking a stranger's camera) were more likely to donate to charity (REGAN, WILLIAMS and SPARLING 1972; CUNNINGHAM, STEINBERG and GREV 1980).

Although the above findings provide direct support for the models that define an apology as unconditional cooperation (i.e., costly prosocial behaviour), mundane apologies (e.g., stating "I am sorry") are neither particularly prosocial nor costly. Nevertheless, OHTSUBO and WATANABE (2009) noted that apologetic remarks, if accompanied by some costly act (e.g., cancelling an important meeting to make an apology as soon as possible), can signal the apologiser's valuation of the endangered relationship to the victim. OHTSUBO and WATANABE empirically tested this hypothesis and found that people perceive costly apologies as more sincere than no-cost apologies (see also OHTSUBO, WATANABE, KIM, KULAS, MULUK, NAZAR, WANG and ZHANG under review, for replications in seven countries). Some researchers have also begun to investigate whether people voluntarily incur some cost to make an apology to their partner in iterated economic games (FISCHBACHER and UTIKAL 2010; HO 2012).

There is another non-prosocial form of costly remorseful display, self-punishment. NELISSEN and ZEELENBERG (2009; NELISSEN 2012) recently demonstrated that participants who had unintentionally inflicted financial damage to a partner voluntarily accepted some financial losses or chose to administer stronger electric shocks to themselves (see also WALLACE and SADALLA 1966; WALLINGTON 1973, for classical demonstrations of self-punishment). Such guilt-induced self-punishment, which NELISSEN and ZEELENBERG

dubbed the Dobby effect, was enhanced by the presence of the victim but not by the presence of someone else (NELISSEN 2012). Accordingly, NELISSEN concluded that guilt-induced self-punishment is primarily targeted to the victim and serves to provide a form of reconciliation.

In sum, the previous studies investigated two human forms of costly remorseful display, costly apology and self-punishment. These studies emphasised reconciliation with a specific partner, and were interested mostly in the direct reciprocity context, demonstrating that people tend to express costly remorseful displays to maintain a relationship with a specific partner. The idea that the primary function of costly remorseful displays is to restore a cooperative relationship with the victim is consistent with the logic of contrite tit-for-tat. We shall refer to this idea as the *relationship maintenance hypothesis*. However, these costly remorseful displays are also useful in the indirect reciprocity context. After erroneously choosing defection, the leading eight strategies express a costly display to restore their public image (or standing). We shall refer to this indirect reciprocity interpretation as the *reputation maintenance hypothesis*.[1]

According to the relationship maintenance hypothesis, transgressors are expected to bear the cost of remorseful displays as far as the cost of losing the endangered relationship exceeds the signalling cost. It is worth noting here that failure to make a costly remorseful display may result in an immediate loss of a valuable relationship. In the indirect reciprocity context, on the other hand, it takes some time for one's bad reputation to be shared by group/societal members. This implies that the remorseful display in the indirect reciprocity context is associated with the commitment problem (FRANK 1988). That is, the loss of good standing is delayed (and thus discounted), while the cost of the display must be borne immediately. Therefore, the evidence for the presence of costly remorseful displays in the direct reciprocity context cannot be readily extended to the indirect reciprocity context.

FRANK (1988) argued that emotions play a pivotal role in solving the commitment problem. If the transgressors feel emotional distress (e.g., guilt, shame) due to their wrongdoing, they will have a non-delayed psychological incentive to incur a cost to alleviate it. As we have seen, previous studies have shown that the sense of guilt induces (1) prosocial behaviour towards both the victim and other people uninvolved in the incident and (2) self-punishment targeted to the victim. Nonetheless, whether the sense of guilt (or some other aversive emotion) also facilitates (a) a costly apology and (b) a private form of self-punishment has not yet been empirically investigated.

1.1. Purposes of the reported experiments

To avoid a questionable research practice (i.e., reporting unexpected findings as predicted; JOHN, LOEWENSTEIN and PRELEC 2012), we explain our original purposes here. The original purpose of Experiment 1 was to provide further evidence for the relationship maintenance hypothesis. Accordingly, it was designed to test the effects of the presence of past or future interactions with a specific partner. The results, however, revealed that the presence of past or future interactions had little effect on the participants' inclinations to make a costly apology. We suspected that the manipulation was not adequate and therefore decided to take a different approach. Experiment 2 was designed to test whether participants would be more willing to incur a cost to express their remorse to a specific partner (i.e., costly apology) than only to themselves (i.e., self-punishment). The relative frequency of self-punishment was, however, almost equivalent to the relative frequency of a costly apology. These results are not congruent with the relationship maintenance hypothesis. Experiment 3, as well as a follow-up study of Experiment 2, was thus conducted as a preliminary test of the reputation maintenance hypothesis.

2. Experiment 1

2.1. Method

2.1.1. Participants, design, and overview of Experiment 1

Participants were 72 Japanese undergraduates (30 males and 42 females, mean age = 18.89 years and range = 18–24 years) who responded to an email invitation that emphasised a 500 Japanese yen (JPY) monetary reward for participation, as well as an extra reward contingent on their performance on an experimental game (500 JPY ≈ 3.50 EUR at the time of experiment). The invitation was sent to the campus-wide participant pool, consisting of those who previously indicated their interest in psychology experiments.

The purpose of Experiment 1 was to examine whether a costly apology would be facilitated by the presence of future interactions with the victim. In addition, the presence of past interactions might foster familiarity with the partner and consequently facilitate a costly apology. Correspondingly, a 2 (past interaction: present vs. absent) × 2 (expectation of future interaction: present vs. absent) between-participants factorial design was employed.[2]

In the present experiment, all participants allocated a fixed amount of money between themselves and their partner. The game was deliberately prepared so that all participants would unintentionally make an unfair allocation decision. Participants were then asked to indicate their willingness to pay some amount of money to send an apology message to their partner. This apology task was inserted into two sessions of the dyadic quiz game whereby the presence of past and future interactions with the allocation game partner was manipulated. Although the first quiz game was played by participants, the second quiz game, which manipulated the *expectation* of future interaction, was not actually played.

The partner arrangements for the four conditions were as follows. In the absent/absent condition (past interaction and future interaction, respectively), participants were told that they would play the first quiz game, the apology task, and the second quiz game with three different persons (Persons A, B, and C). In the absent/present condition, the

partner for the first quiz game was Person A and the partner for the apology task and the second quiz game was Person B. These arrangements are denoted as ABC and ABB, respectively. Accordingly, the remaining present/absent and present/present conditions can be denoted as AAB and AAA, respectively. In the absent/absent condition, three partners (Persons A, B, and C) were needed for each participant to be paired with a different partner in each of the three games. Therefore, each experimental session involved four participants. Partitions were placed among them so that they were unable to see each other. When there were mutual friends in a single session, it was arranged so that mutual friends would not be paired in any of the three games.

2.1.2. Quiz game

In the quiz game, a pair of participants was asked to guess target words collaboratively. For each target word, one of the participants served as a guesser and the other a cue-giver. The cue-giver sent a cue (i.e., a word related to the target word) to the guesser. The guesser guessed the target word using the cue. If the guess was incorrect, the cue-giver sent a second cue and the guesser made another guess. These procedures were repeated until the guesser made a correct guess. Once the guesser made a correct guess, a new target word was given to the pair. The quiz game was played for 10 minutes. After five minutes elapsed, the guesser and cue-giver roles were switched within the pair.

2.1.3. Apology task

Whether participants would make a costly apology or not was determined by having participants behave in an unfair manner. After the first quiz game, participants received instructions regarding the so-called "resource allocation game": Each participant, paired with another participant, would allocate a sum of 500 JPY between himself/herself and a partner. The pairs would simultaneously engage in the same task and each participant's monetary reward for the game would be "the amount he/she reserved for him-/herself" plus "the

amount the partner gave to him/her." This mutual allocation procedure was employed to minimize the concern for outcome equality (FEHR and SCHMIDT 1999; DAWES, FOWLER, JOHNSON, MCELREATH and SMIRNOV 2007). In making the apology decision, participants were not told about the amount of the inequality that would result from their allocation decisions.

Participants were then told that there were two experimental conditions, referred to as the "intention condition" and the "no intention condition." All participants were in fact assigned to the no intention condition in an ostensibly random manner. Participants were told that their partner would not know which condition they had been assigned to. Therefore, the intention of their later unfair allocation would remain ambiguous to their partner. According to FISCHBACHER and UTIKAL (2010), without such ambiguity in the transgressor's intention, apologies would not facilitate the victim's forgiveness.

Participants were told that in the no intention condition they would draw one card from nine cards placed face down on a table. It was explained that each of the nine cards represented one of nine resource allocation schemes: (self, partner) = (50, 450), (100, 400) … (400, 100), (450, 50). However, all the cards carried the unfair allocation scheme of (400, 100). After drawing a card and confirming that they would give 400 JPY to themselves and only 100 JPY to the partner, participants filled out a post-task questionnaire. This questionnaire included a guilt item (i.e. "How guilty are you feeling about the allocation?" accompanied by a 5-point scale: 1 = "not at all" – 5 = "very strongly").[3] The Japanese word *zaiakukan* was used for the English word *guilt* in this item.

After completing the post-task questionnaire, participants were told that those who had allocated more than half to themselves would be allowed to send an apology message to the partner. Participants first indicated their willingness to send the message ("yes" or "no"). For those who answered "yes" (henceforth referred to as *costly apologisers*), their *reservation*

price for apology was assessed as follows. To send the message, participants would have to pay a pre-determined amount (x JPY) to the experimenter. Not knowing the exact value of x, participants indicated the maximum cost they would be willing to incur (y JPY). If $y \geq x$, participants would be allowed to send the message and charged x JPY. On the other hand, if $y < x$, participants would pay nothing and would not be allowed to send the message. This procedure was employed because based on the auction theory (STEIGLITZ 2007), it was considered that participants' honest report of their reservation price would be facilitated by not requiring them to pay the reported reservation price, y JPY.

2.1.4. Debriefing

After the apology task, each participant was thoroughly debriefed.[4] The debriefing included explanations of why the deceptive procedures were required. All participants then received 1000 JPY, which was equivalent to the monetary reward when the equal allocation (i.e., 250 JPY for each) was made by both players: 250 JPY from one's own allocation plus 250 JPY from the partner's allocation plus an additional guaranteed 500 JPY for participation.

2.2. Results and discussion

Among the 72 participants, 45 participants (62.5%) were willing to send the apology message (i.e., costly apologisers). Their reservation prices ranged from 50 to 300 JPY. The reservation price data (including 0 JPY assigned to the non-apologisers) were submitted to a 2 (past interaction) × 2 (future interaction) between-participants analysis of variance (ANOVA). Neither the main effects nor the interaction effect was significant, all $F(1, 68)$'s < 1.00: mean±SD reservation prices were 93.75±91.06 (the absent/absent condition), 67.50±71.22 (the absent/present condition), 75.00±81.91 (the present/absent condition), and 85.70±69.52 (the present/present condition). These effects were also non-significant after excluding the non-apologiser data. Consistent with these results, the relative frequency of the

costly apologisers did not significantly differ across the conditions, .69 (absent/absent), .55 (absent/present), .60 (present/absent), and .69 (present/present): all $|Z|$'s < 1 by a logistic regression analysis. In addition, no significant effects emerged from a 2 × 2 ANOVA with self-reported guilt as a dependent variable, all $F(1, 68)$'s < 1.00: mean±SD guilt scores were 3.50±1.21 (absent/absent), 3.55±1.19 (absent/present), 3.40±1.10 (present/absent), and 3.44±1.09 (present/present).

Despite the lack of significant effects of past and future interactions on the two primary dependent variables, the reservation prices were highly correlated with self-reported guilt: $r_{72} = .49$, $p < .001$, where the subscript attached to r represents sample size (*Figure 1a*). This pattern was consistent in the four conditions, although two of the four product-moment correlations did not reach the conventional significance level due to small sample sizes: $r_{16} = .66$, $p = .005$ (absent/absent); $r_{20} = .35$, $p = .135$ (absent/present); $r_{20} = .53$, $p = .017$ (present/absent); $r_{18} = .43$, $p = .100$ (present/present). Consistent with the correlation analyses, the costly apologisers (3.91±0.87) reported a stronger sense of guilt than did the non-costly apologisers (2.74±1.13), $t_{70} = 4.92$, $p < .001$.

The significant correlation between the costly apology and self-reported guilt indicates that the observed costly apologies were not made on a whim. Nonetheless, the effects of the past and future interactions were not significant. Therefore, Experiment 1 failed to support the relationship maintenance hypothesis, which predicts that the possibility of a future interaction facilitates costly apology-making. In considering this result, we note that it was possible that the experimental manipulations were not sufficiently strong given that no monetary incentives were involved in the two sessions of the quiz game. Monetary incentives in the quiz game might encourage some participants to incur a cost for apology in order to earn as much money as possible in the second quiz game (Ho, 2012). However, such a calculation-based costly apology is not equivalent to the current subject of interest (i.e.,

emotionally driven costly displays). We thus decided to take a different approach to test the relationship maintenance hypothesis. In particular, we reasoned that if the costly remorse display is targeted at the specific partner, the willingness to incur a cost would be substantially curtailed by explaining that the partner would not see their remorseful display. Therefore, it was predicted that an anonymous form of self-punishment would be observed less frequently than a costly apology.

3. Experiment 2

3.1. Method

In Experiment 2, the modes of remorse display (apology vs. self-punishment) were manipulated as a between-participants factor. The apology condition was conducted following procedures identical to those of Experiment 1 with two exceptions. First, the quiz games were not included. Second, participants were led to believe that their partner was in another room and were tested individually. In this section, we shall avoid duplicating the explanations of the apology condition.[5]

3.1.1. Participants and design

Participants were 61 Japanese undergraduates (29 males and 32 females; mean age = 18.8 years and range = 18–25 years) who were recruited in the same manner as in Experiment 1. Participants were randomly assigned to either the apology or the self-punishment condition. These two conditions differed only in the questionnaire that assessed participants' willingness to incur the cost.

3.1.2. Self-punishment condition

In the self-punishment condition, the participants were told that they would be allowed to reduce their monetary reward if they were dissatisfied with the allocation they had made. Those who were willing to do this (referred to as "self-punishers") were further asked

to indicate the maximum amount of money they were willing to lose (y JPY). The instructions read that there was a predetermined amount x JPY, which would be deducted from the reward if $y \geq x$, while nothing would happen if $y < x$. To ensure that participants understood the private nature of the self-punishment, the instructions emphasized that the deducted amount would not be transferred to the partner. The instructions also emphasised that the partner would not be informed of whether they decided to reduce their reward. The experimenter used neutral expressions and avoided using any words connoting "punishment."

One might have some concern about the issue of mundane realism. People would not throw their money away after committing transgressions in real life, as no one would inflict electric shock on themselves. Nonetheless, either operationalisation conceptually corresponds to the notion of self-punishment. More importantly, the present operationalisation of self-punishment allowed us to assess both self-punishment and costly apology in a comparable manner.

3.2. Results and discussion

Contrary to the prediction from the relationship maintenance hypothesis, the relative frequency of a costly apology (.52 = 16/31) was almost equivalent to that of self-punishment (.47 = 14/30), $p = .80$ by Fisher's exact test. A t-test including all participants indicated that the reservation price did not significantly differ between the two conditions (those who were not willing to incur any cost were assigned 0 JPY as their reservation price): mean±SD = 77.42±84.50 and 53.33±62.88 for the apology and self-punishment conditions, respectively, $t_{59} = 1.26$, *ns* (a non-parametric test also confirmed the non-significant difference between the conditions). When only apologisers and self-punishers were included in the analysis, the mean reservation prices were significantly higher in the apology condition, 150±51.64, than the self-punishment condition, 114.29±36.31, $t_{28} = 2.16$, $p = .039$. A non-parametric test further confirmed the significant difference, $p = .027$.

The non-significant difference in the frequency of the two types of signallers suggests that whether or not one was willing to incur the cost was not affected by the type of signal (costly apology vs. self-punishment). Consistent with this interpretation, the product-moment correlation coefficients between the reservation price and self-reported guilt were significant in both conditions, $r_{31} = .65$, $p < .001$ (*Figure 1b*) and $r_{30} = .47$, $p = .008$ (*Figure 1c*) in the apology and self-punishment conditions, respectively. That is, the same proximate emotion elicited both costly apologies and self-punishment.

3.3. Follow-up Study

Given that the results are inconsistent with the relationship maintenance hypothesis, we turned to the reputation maintenance hypothesis and the notion of emotional commitment (FRANK 1988). If those who are emotionally committed to fairness, and specifically *egalitarianism* given the task of the present experiment, feel a stronger sense of guilt and are more likely to resort to a costly remorseful display, it serves the signallers to maintain their reputation as an egalitarian. Since we observed a large individual difference in Experiment 2 (i.e., about half the participants were costly signaller and the other half non-signallers), it was appropriate to test whether the costly signallers were endorsers of egalitarianism.

3.3.1. Method of the follow-up study

We invited the participants in Experiment 2 to a follow-up study via email. Of the 61 participants in the original study, 32 (15 from the apology condition and 17 from the self-punishment condition) took part in the follow-up study. Participants individually completed a questionnaire involving OHTSUBO, KAMEDA and KIMURA's (1996) measure of preferences for various resource distribution principles.

Using OHTSUBO et al.'s (1996) measure, we asked participants their opinions regarding a hypothetical resource allocation situation. They were presented with a hypothetical scenario in which three group members contributed 1, 2, and 3 units to a group

achievement, respectively. Given the ratio of contributions, participants rated the desirability of six allocation schemes with a 201-point scale (–100 = "absolutely undesirable" to +100 = "absolutely desirable"). The six hypothetical allocation schemes for the high, middle, and low contributors, respectively, were the following: (6, 4, 2), (4, 4, 4), (2, 4, 6), (18, 12, 6), (12, 12, 12), and (6, 12, 18). The second and fifth schemes were in agreement with the equality principle. The first and fourth schemes represented the equity principle, which prescribes the allocation of resources according to each member's contribution (ADAMS 1965). The third and sixth schemes were completely unfair in the sense that they were neither equal nor equitable. However, the sixth allocation scheme was Pareto-efficient over the first (equitable) and second (equal) schemes (i.e., the sixth scheme gave every member a larger reward than the first and second schemes except that it gave the equal reward to the first member as the first scheme). The six desirability ratings were standardised within each participant and the standardised desirability ratings for the two equal schemes were averaged. This egalitarian score thus reflected each participant's relative endorsement of the equality principle over the equity principle and over efficiency.

3.3.2. Results of the follow-up study

Due to the small sample size, the two conditions in the original experiments were collapsed. The product-moment correlation between the reservation price and the egalitarian score controlling for the condition in the original experiment was not statistically significant, $r_{32} = .21$, *ns*. However, in the apology condition, there was an outlier participant whose reservation price (300 JPY) deviated 2.93 SD from the mean (see *Figure 1b*). Once this participant was excluded, the correlation was marginally significant, $r_{31} = .34$, $p = .060$.[6]

When the two conditions were analysed separately, the product-moment correlation between the reservation price and the egalitarian score was higher in the self-punishment condition, $r_{17} = .49$, $p = .048$, than in the apology condition, $r_{14} = .22$, $p = .460$ (the difference

between the two correlation coefficients was not significant, $Z = .78$). As we have noted, the reservation price was higher in the apology condition than the self-punishment condition. This implies that the reservation price in the apology condition reflected both reputational and interpersonal concerns, and thus the correlation in the apology condition might have been diminished.

The results of the follow-up study were at least consistent with the prediction from the reputation maintenance hypothesis. Nevertheless, the follow-up study had some shortcomings (e.g., a relatively large time lag between the original study and the follow-up study, small sample size). In Experiment 3, we therefore sought to replicate the self-punishment condition of Experiment 2 as well as the follow-up study.

4. Experiment 3

The primary purpose of Experiment 3 was to replicate the positive correlation between egalitarian attitude and reservation price for self-punishment. In addition, Experiment 3 included a different measure of guilt to validate the single item measure used in Experiments 1 and 2.

4.1. Method

Participants were 43 Japanese undergraduates (24 males and 19 females; mean age = 19.1 years and range = 18–21 years) who were recruited in the same manner as in Experiments 1 and 2. One male participant who suspected the use of deceptive procedures was discarded from the analyses.

As in the self-punishment condition in Experiment 2, participants engaged in the resource allocation task, rated their sense of guilt, and reported their willingness to inflict self-punishment. To minimise the participants' apprehension of evaluation by the experimenter, participants were told that their self-punishment decision would not be

assessed by the experimenter who would directly interact with the participants. Another experimenter who would not see the participants would examine their self-punishment decision and prepare their monetary reward. The interacting experimenter would bring the enveloped reward to the participants.

In addition to the single-item measure of guilt, the State Shame and Guilt Scale (TANGNEY and DEARING 2002) was administered. This scale was designed to assess a participant's current feelings of shame, guilt, and pride using five items each. The items included "I feel worthless, powerless" (shame), "I feel remorse, regret" (guilt), and "I feel worthwhile, valuable" (pride). Cronbach's α coefficients were .79, .76, and .75, for the shame, guilt, and pride subscales, respectively. After the self-punishment task, participants received a questionnaire including OHTSUBO et al.'s (1996) egalitarianism measure. After completing the questionnaire, participants were fully debriefed and paid 1000 JPY.

4.2. Results and Discussion

We first confirmed that the results of Experiment 3 were comparable to those of the self-punishment condition in Experiment 2. Approximately half of the participants (.52 = 22/42) were willing to inflict self-punishment. This was not significantly different from the relative frequency of self-punishment in Experiment 2 (.47 = 14/30), $\chi^2_1 = .03$. The mean±SD reservation price was 139.00±53.79, which was slightly (but not significantly) greater than the mean reservation price in the self-punishment condition of Experiment 2, $t_{32} = 1.49$, $p = .07$. The marginally significant difference was possibly due to an outlier participant whose reservation price was 300 JPY (see *Figures 1c* and *1d*). Therefore, despite the procedure introduced in Experiment 3 so as to minimise participants' apprehension of evaluation by the experimenter, neither the frequency of self-punishment nor the reservation price for it decreased.

Self-reported guilt was significantly correlated with the reservation price, $r_{42} = .42$, p

= .005. In Experiment 3, the state guilt, shame, and pride scores were obtained by using the State Shame and Guilt Scale. The single-item guilt score was significantly correlated with the state guilt score, $r_{42} = .50, p < .001$. However, the state guilt score was not significantly correlated with the reservation price, $r_{42} = .22, p = .157$. Notice that the single-item guilt measure probed guilty-feelings, explicitly referring to the unfair allocation, while the state guilt items measured current feeling without any explicit referent. This might account for the lack of a significant correlation with the reservation price. Parenthetically, the single-item guilt score was also significantly correlated with the state shame score, $r_{42} = .58, p < .001$, and the state pride score, $r_{42} = -.43, p = .005$. Moreover, among the three self-conscious emotions measured by TANGNEY and DEARING's scale, only the state pride score was significantly correlated with the reservation price, $r_{41} = -.48, p = .002$. We shall consider a possible interpretation of this result in the General Discussion section.

More central to the purpose of Experiment 3, the product-moment correlation between the egalitarian score and the reservation price for self-punishment was significant, $r_{42} = .52, p < .001$ (*Figure 2*). This result confirmed the follow-up study of Experiment 2. Interestingly, however, the egalitarian score was not significantly correlated with the self-reported guilt, $r_{42} = .15, p = .351$. When entered into a multiple regression analysis, both self-reported guilt ($\beta = .35, p = .007$) and the egalitarian score ($\beta = .47, p < .001$) were found to be significant predictors of the reservation price, $F_{2, 39} = 12.74, p < .001$, adjusted $R^2 = .36$ for the overall model. This result suggests that emotion and egalitarian attitude (or cognition) could function as distinct elicitors of self-punishment.

5. General Discussion

The set of three experiments indicated that costly remorseful displays after an unintentional transgression were not fully explained by the relationship maintenance

hypothesis alone. The pattern was also consistent with the reputation maintenance hypothesis. In Experiment 1, the presence of past or future interactions did not affect the reservation price for a costly apology. In Experiment 2, whether the costly signals would be witnessed by the victim did not affect the frequency of the signals: the relative frequency of self-punishment was almost equal to that of a costly apology. These results are not readily explained if we only consider that the costly remorseful displays are targeted at a specific interaction partner. We then tested a prediction from the reputation maintenance hypothesis: egalitarians (who should be more concerned about their reputation as egalitarians) were more prone to show the costly remorseful displays after the violation of the equality norm. Experiment 3, as well as the follow-up of Experiment 2, confirmed this prediction. Notice, however, that this prediction is based on an implicit assumption that the presence of an audience (or rational calculation taking the audience effect on reputation into account) is not necessary for the costly displays to be elicited. Perhaps those who are concerned with their reputation are emotionally bound to produce costly signals. There is in fact some evidence that reputational concern facilitates prosocial behaviour outside rational reasoning (e.g., HALEY and FESSLER 2005). However, participants might have apprehended the evaluation by the experimenters. Although we tried to minimise this apprehension in Experiment 3, the apprehension reduction procedure made no marked change in participants' decision to inflict self-punishment.

The findings above seem to have some theoretical implications for the indirect reciprocity literature. As we have noted earlier, the typical model of indirect reciprocity assumes that an unintentional non-cooperator will express an "apology" by unconditional cooperation, which implies accepting punishment (i.e., defection) by another player. Under this mechanism, however, the punisher might obtain a poor standing, which can have a detrimental effect on the stability of cooperation by triggering a chain of punishments on previous punishers. It is known that such a spiral of unnecessary punishments can be avoided

if the punisher's next partner is capable of discriminating justified defection from unjustified defection (NOWAK and SIGMUND 2005; OHTSUKI and IWASA 2006). Such a discriminating strategy, however, might be cognitively taxing even for human players (MILINSKI et al. 2001). If unintentional non-cooperators can immediately give a costly signal of their benign intention and will be forgiven, the detrimental effect of implementation errors may be substantially lessened. It seems worthwhile to empirically investigate whether the costly remorseful displays would facilitate cooperation under noisy indirect reciprocity environments.

The present study indicated several unresolved issues that should be investigated in future studies. First, the result of Experiment 2 appears contradictory to that of NELISSEN (2012). In NELISSEN's experiment, the presence of the victim facilitated the self-punishment but the presence of general audience did not. Therefore, NELISSEN's result supported the relationship maintenance hypothesis. Although the presence of the victim did not affect the frequency of costly signals in Experiment 2 (i.e., the relative frequency of self-punishment was not significantly different from that of costly apology), the reservation prices were greater in the self-punishment condition than the apology condition. Thus Experiment 2 also provided partial support for the relationship maintenance hypothesis. As the reservation price, a continuous measure of self-punishment, is closer to the measure of self-punishment of NELISSEN's experiment (i.e., the acceptable level of electric shocks), the result of Experiment 2 is considered rather similar to NELISSEN's result. The apparent contradiction was due to the dichotomous measure of self-punishment, which was uniquely employed in the present study.

Second, the reservation price in Experiment 3 was independently predicted by self-reported guilt and egalitarian scores. A straightforward prediction from the reputation maintenance hypothesis may be that egalitarians, who are more concerned for their reputation, would feel a stronger sense of guilt for their violation of the relevant norm. This sense of guilt

would then elicit the costly remorseful display. However, the observed pattern does not support this prediction. Further studies are therefore needed to explore the relation among attitudes, emotions, and behaviour in the context of costly remorseful displays.

Finally, the result was not decisive regarding the nature of the emotion that elicited the costly remorseful displays. Although the Japanese word, *zaiakukan*, is commonly translated to the English word, guilt, self-reported *zaikakukan* was significantly correlated with the three types of self-conscious emotions—guilt, shame, and pride. This is not surprising given that guilt and shame are often experienced together and people are typically unclear about the distinction (TANGNEY and DEARING 2002). Another curious but puzzling finding was that only the pride score was significantly (negatively) correlated with the reservation price. One possible interpretation is that the pride items tapped a subtle change in the feeling of shame. Notice that some pride items (e.g., I feel worthwhile, valuable) are almost reversed versions of the shame items (e.g., I feel worthless, powerless). Although the self-conscious emotion literature tends to implicate guilt, but not shame, as a cause of prosocial behaviour (TANGNEY and DEARING 2002; but see DE HOOGE, BREUGELMANS and ZEELENBERG 2008), shame possibly played a more important role in the present study, which examined post-transgression reactions (KELTNER, YOUNG and BUSWELL 1997; FESSLER 2007; GINER-SOROLLA, CASTANO, ESPINOSA and BROWN 2008). This interpretation that shame, rather than guilt, was evoked by the manipulation of our experiments is also consistent with BAUMEISTER, STILLWELL and HEATHERTON's review (1994) revealing that guilt functions to enhance relationships (i.e., the function presumed by the relationship maintenance hypothesis). Perhaps, the present study addressed a shame-related phenomenon, and thus one not directly relevant to guilt-related (relationship enhancing) phenomena. Future research employing more refined measures of self-conscious emotions is needed to fully understand the emotional underpinnings of costly apology and self-punishment.

In sum, the present research demonstrated that people not only make a costly apology but also inflict self-punishment after committing an unintentional transgression. The observed self-punishment is difficult to explain if one restricts the function of costly remorseful display to relationship maintenance. A plausible alternative function of self-punishment is to maintain one's reputation. The present study provided some support for this interpretation by showing that those who are supposed to care more about their egalitarian reputation are more likely to inflict self-punishment after splitting a resource unequally. The two separate functions seem to correspond to the two types of apology: private apology and public apology. Thus, this line of research may shed some light on the problem of what factors or emotions encourage people to apologise to whom.

Acknowledgment

The authors are grateful to Hisashi Ohtsuki for his helpful comments on an earlier draft. This research was supported by the Japan Society for the Promotion of Science (No. 21683006).

Footnotes

[1] Although it has been well established that people are concerned with their reputation and that reputational concern makes people more cooperative (e.g., BARCLAY and WILLER 2007; BERECZKEI, BIRKAS and KEREKES 2010; MILINSKI et al. 2001), how this reputational concern facilitates remorseful displays after unwittingly committing a transgression is yet to be investigated.

[2] Experiment 1 did not include a no-transgression control condition, which was included in the previous studies (e.g., REGAN et al. 1972; CUNNINGHAM et al. 1980). The dependent variable of those studies was prosocial behaviour unrelated to the transgression. Therefore, it

was necessary for the researchers to assess the base rate of the prosocial behaviour. In Experiment 1, on the other hand, the dependent variable was the willingness to make a costly apology. It is weird to ask participants who did not commit any transgression their willingness to apologise. Accordingly, we did not include the no-transgression control condition. However, as a validity check, we correlated the willingness to apologise with the reported level of guilt.

[3] Although we do not report other items included in the post-allocation questionnaire (see SIMMONS, NELSON and SIMONSOHN 2011, for possible criticisms), the information of those items is available from the corresponding author upon request. However, the only item that was significantly correlated with the costly apology/self-punishment throughout the three experiments was the guilt item reported in the main text.

[4] We were cognizant of the ethical issues of deceptive procedures and their detrimental effects on the participant pool (ORTMANN and HERTWIG 2002). However, deceptive procedures were required here for the following reasons. (i) The present experiment theoretically implicated emotional responses to one's unfair behaviour as a cause of a costly apology. The strategy method, a standard method to avoid deceptive procedures in economic experiments, tends to attenuate emotional responses (CASARI and CASON 2009; COOK and YAMAGISHI 2008). (ii) The deceptive procedures facilitated efficiency of the data collection (COOK and YAMAGISHI 2008). In our previous study involving a similar resource allocation task, the majority of the participants divided the resource nearly equally. Without the deceptive procedures, therefore, a large portion of participants would need to be discarded from the data set. (iii) The deceptive procedures ensured the experimental control (BONETTI 1998). It is reasonable to expect participants who make a fair allocation decision to be more inclined to make a costly

apology. The presence of such naturally confounded variables would undermine the experimental control. (iv) The inclusion of the second quiz game would force participants to stay in the laboratory for an unnecessarily long time.

[5] Experiments 2 and 3, as well as the follow-up of Experiment 2, included several individual difference measures that are not reported in the main text. The information of these measures is available from the corresponding author upon request.

[6] One of the reviewers of this article pointed out that those who endorsed the equity principle might also inflict self-punishment because the only conceivable equitable allocation in the experimental setting, in which neither player contributed to the acquisition of the allocated resource, was the equal allocation. We examined the correlation between the reservation price for costly displays and the equity score, which was computed in the similar manner as the egalitarian score so that it reflected each participant's relative endorsement of the equity principle over the equality principle and over efficiency. By definition, it was negatively correlated with the egalitarian score, $r_{32} = -.92$, $p < .001$. Accordingly, the equity score was negatively, but not significantly, correlated with the reservation price for costly displays, $r_{32} = -.11$ (or $-.24$ after excluding the outlier data point) both *ns*. The same pattern held for Experiment 3. The equity score was negatively correlated with the egalitarian score, $r_{42} = -.85$, $p < .001$, and the reservation price for self-punishment, $r_{42} = -.44$, $p = .003$, respectively. This pattern strongly depends on OHTSUBO et al.'s operational definition of the egalitarian attitude. For future research, it is worth testing whether the self-punishment tendency would be correlated with some other social preferences, for example, those operationalised as the social value orientations (VAN LANGE, OTTEN, DE BRUIN and JOIREMAN 1997).

References

ADAMS, J. S. (1965): Inequity in social exchange. In: Berkowitz, L. (ed.): *Advances in Experimental Social Psychology* (Vol. 2). New York: Academic Press, pp. 267-299.

BARCLAY, P. and WILLER, R. (2007): Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B*, 274, 749-753.

BAUMEISTER, R. F., STILLWELL, A. M. and HEATHERTON, T. F. (1994): Guilt: An interpersonal approach. *Psychological Bulletin*, 115, 243-267.

BERECZKEI, T., BIRKAS, B. and KEREKES, Z. (2010): Altruism towards strangers in need: Costly signaling in an industrial society. *Evolution and Human Behavior*, 31, 95-103.

BONETTI, S. (1998): Experimental economics and deception *Journal of Economic Psychology*, 19, 377-395.

BOYD, R. (1989): Mistakes allow evolutionary stability in the repeated prisoner's dilemma game. *Journal of Theoretical Biology*, 136, 47-56.

CASARI, M. and CASON, T. N. (2009): The strategy method lowers measured trustworthy behavior. *Economics Letters*, 103, 157-159.

COOK, K. S. and YAMAGISHI, T. (2008): A defense of deception on scientific grounds. *Social Psychology Quarterly*, 71, 215-221.

CUNNINGHAM, M. R., STEINBERG, J. and GREV, R. (1980): Wanting to and having to help: separate motivations for positive mood and guilt-induced helping. *Journal of Personality and Social Psychology*, 38, 181-192.

DAWES, C. T., FOWLER, J. H., JOHNSON, T., MCELREATH, R. and SMIRNOV, O. (2007): Egalitarian motives in humans. *Nature*, 446, 794-796.

DE HOOGE, I. E., BREUGELMANS, S. M. and ZEELENBERG, M. (2008): Not so ugly after all: when shame acts as a commitment device. *Journal of Personality and Social*

Psychology, 95, 933-943.

FEHR, E. and SCHMIDT, K. M. (1999): A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114, 817-868.

FESSLER, D. M. T. (2007): From appeasement to conformity: evolutionary and cultural perspectives on shame, competition, and cooperation. In Tracy, J. L., Robins, R. W., Tangney. J. P. (eds.): *The Self-Conscious Emotions: Theory and Research*. New York: Guilford, pp. 174-193.

FISCHBACHER, U. and UKITAL, V. (2010): On the acceptance of apologies. Research Paper Series Thurgau Institute of Economics and Department of Economics at the University of Konstanz

FRANK, R. H. (1988): *Passion within Reason: The Strategic Role of Emotions.* New York: Norton.

GINER-SOROLLA, R., CASTANO, E., ESPINOSA, P. and BROWN, R. (2008): Shame expressions reduce the recipient's insult from outgroup reparations. *Journal of Experimental Social Psychology*, 44, 519-526.

HALEY, K. J. and FESSLER, D. M. T. (2005): Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, 26, 245-256.

HO, B. (2012): Apologies as signals: With evidence from a trust game. *Management Science*, 58, 141-158.

JOHN, L. K., LOEWENSTEIN, G. and PRELECT, D. (2012): Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, 23, 524-532.

KELTNER, D., YOUNG, R. C. and BUSWELL, B. N. (1997): Appeasement in human emotion, social practice, and personality. *Aggressive Behavior*, 23, 359-374.

KETELAAR, T. and AU, W. T. (2003): The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: an affect-as-information interpretation of the role of emotion in social interaction. *Cognition and Emotion*, 17, 429-453.

MILINSKI, M., SEMMANN, D., BAKKER, T. C. M. and KRAMBECK, H-. J. (2001): Cooperation through indirect reciprocity: image scoring or standing strategy? *Proceedings of the Royal Society of London B*, 268, 2495-2501.

NELISSEN, R. M. A. (2012): Guilt-induced self-punishment as a sign of remorse. *Social Psychological and Personality Science*, 3, 139-144.

NELISSEN, R. M. A. and ZEELENBERG, M. (2009): When guilt evokes self-punishment: evidence for the existence of a Dobby effect. *Emotion*, 9, 118-122.

NOWAK, M. A. and SIGMUND, K. (1998): Evolution of indirect reciprocity by image scoring. *Nature*, 393, 573-577.

NOWAK, M. A. and SIGMUND, K. (2005): Evolution of indirect reciprocity. *Nature*, 437, 1291-1298.

OHTSUBO, Y., KAMEDA, T. and KIMURA, Y. (1996): When a sense of justice hinders social efficiency: Pareto axiom revisited. *Japanese Journal of Psychology*, 67, 367-374.

OHTSUBO, Y. and WATANABE, E. (2009): Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior*, 30, 114-123.

OHTSUBO, Y., WATANABE, E., KIM, J., KULAS, J. T., MULUK, H., NAZAR, G., WANG, F. and ZHANG, J. (2012): *Are costly apologies universally perceived as being sincere?: A test of the costly apology-perceived sincerity relationship in seven countries*. Manuscript submitted for publication.

OHTSUKI, H. and IWASA, Y. (2006): The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, 239, 435-444.

ORTMANN, A. and HERTWIG, R. (2002): The costs of deception: Evidence from psychology. *Experimental Economics*, 5, 111-131.

REGAN, D. T., WILLIAMS, M. and SPARLING, S. (1972): Voluntary expiation of guilt: A field experiment. *Journal of Personality and Social Psychology*, 24, 42-45.

SIMMONS, J. P., NELSON, L. D. and SIMONSOHN, U. (2011): False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359-1366.

STEIGLITZ, K. (2007): *Snipers, Shills, and Sharks: eBay and Human Behavior*. Princeton, NJ: Princeton University Press.

SUGDEN, R. (1986): *The Economics of Rights, Co-operation and Welfare*. Oxford, UK: Blackwell.

TANGNEY, J. P. and DEARING, R. L. (2002): *Shame and Guilt*. New York: Guilford.

VAN LANGE, P. A. M., OTTEN, W., DE BRUIN, E. M. N. and JOIREMAN, J. A. (1997): Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. *Journal of Personality and Social Psychology*, 73, 733-746.

WALLACE, J. and SADALLA, E. (1966): Behavioral consequences of transgression: I. The effects of social recognition. *Journal of Experimental Research in Personality*, 1, 187-1994.

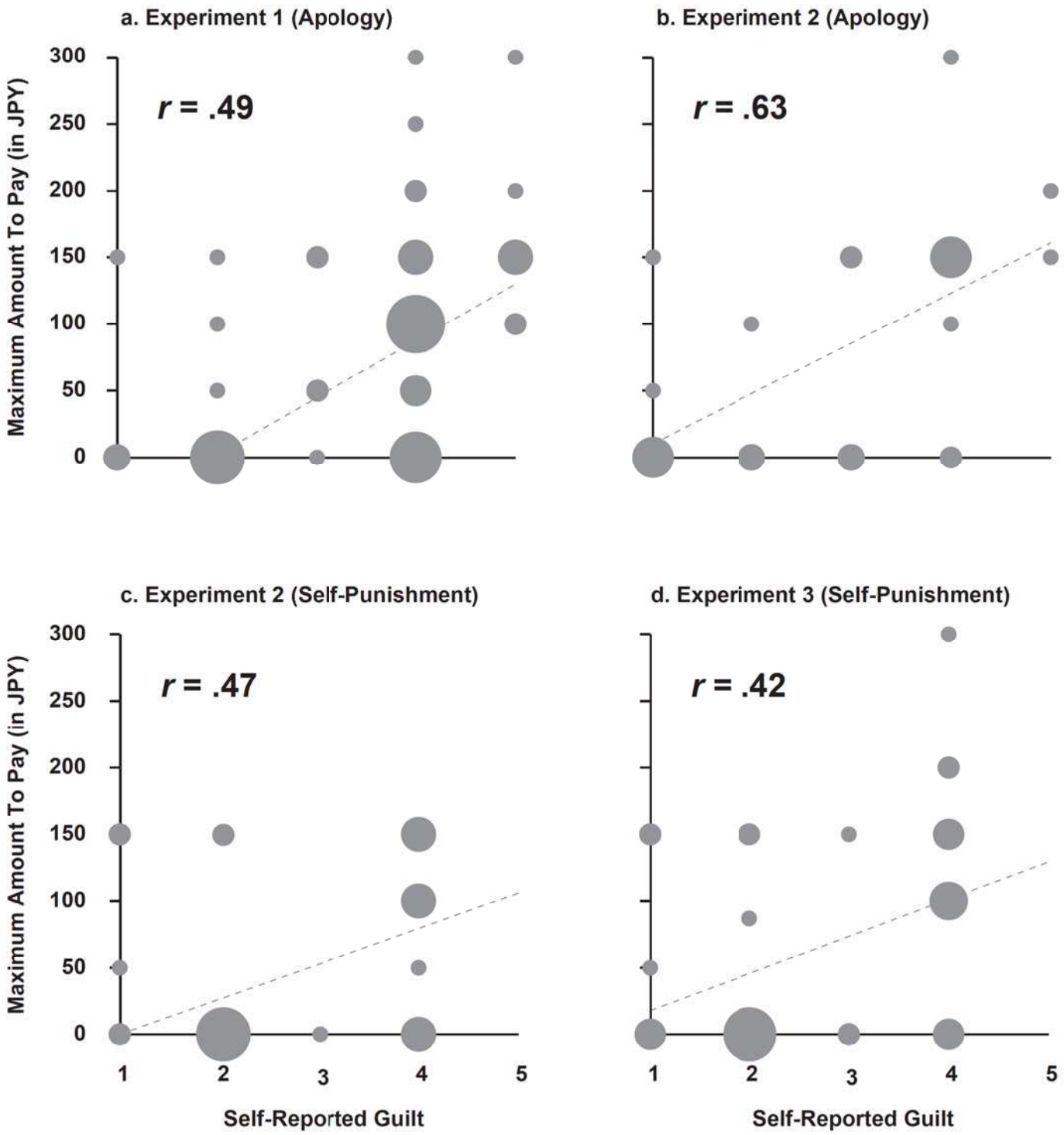
WALLINGTON, S. A. (1973): Consequences of transgression: self-punishment and depression. *Journal of Personality and Social Psychology*, 28, 1-7.

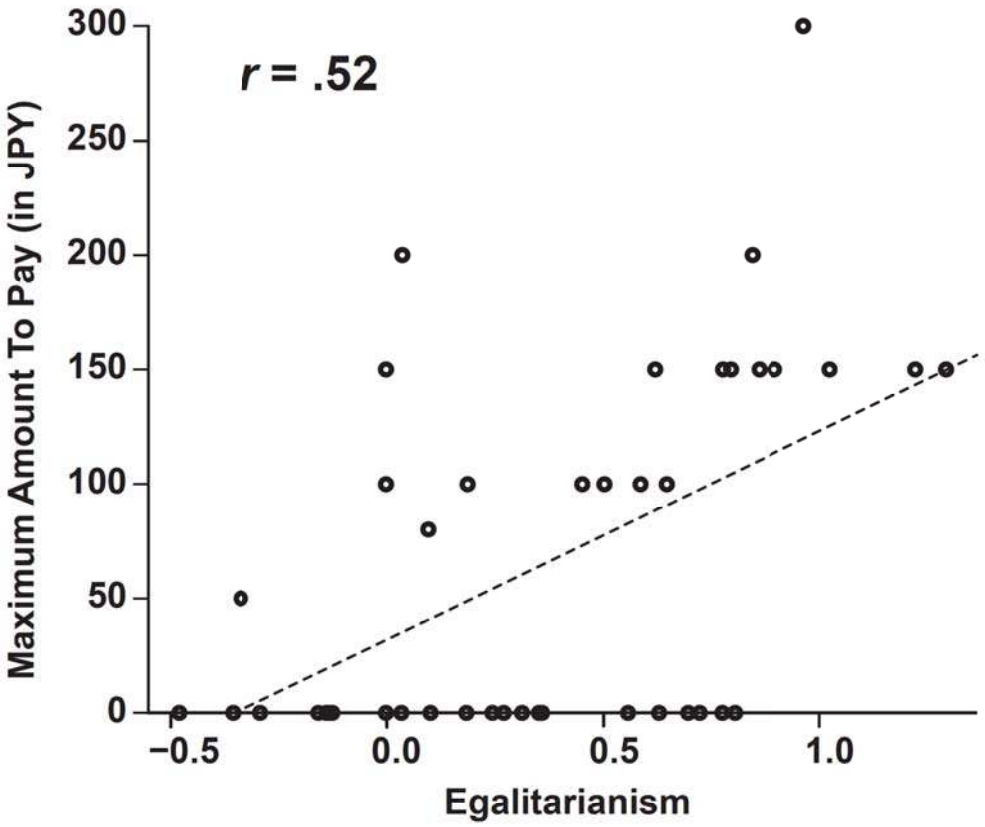
WU, J. and AXELROD, R. (1995): How to cope with noise in the iterated prisoner's dilemma. *Journal of Conflict Resolution*, 39, 183-189.

Figure Captions

Figure 1. Bubble charts showing the distribution of the maximum amount to pay for the costly signals as a function of self-reported guilt: (a) Experiment 1, (b) the apology condition of Experiment 2, (c) the self-punishment condition of Experiment 2, and (d) Experiment 3. The radius of each bubble is proportional to the square root of the frequency at each data point, so that the area represents the frequency. The broken lines are the least-squares regression lines.

Figure 2. Scatter plot showing the relation between egalitarian scores and the maximum amount to pay for self-punishment (Experiment 3). The broken line is the least-squares regression line.


a. Experiment 1 (Apology)
r = .49
b. Experiment 2 (Apology)
r = .63
c. Experiment 2 (Self-Punishment)
r = .47
d. Experiment 3 (Self-Punishment)
r = .42
Maximum Amount To Pay (in JPY)
300
250
200
150
100
50
0
1
2
3
4
5
Self-Reported Guilt


r = .52
Maximum Amount To Pay (in JPY)
300
250
200
150
100
50
0
−0.5
0.0
0.5
1.0
Egalitarianism