Saving face through preference signaling and obligation avoidance

Saving Face Through Preference Signaling and Obligation Avoidance

Matthew Chao*
Williams College

Jonathan Chapman^
NYU Abu Dhabi

Abstract:

Many individuals act more selfishly in games when actions are hidden and their image is not at risk. However, some individuals may still desire to publicly signal reciprocity or other socially desired behavior in these contexts. These individuals may view hidden actions not as an opportunity to act selfishly, but rather as an obstacle to signaling preferences or type. Study 1 tests this by implementing a trust game where nature stochastically intervenes and allocates nothing in place of the second-mover's choice. When nature intervenes, many second-movers choose to sacrifice pay in order to truthfully signal that they attempted to allocate more, and that they therefore tried to reciprocate. Since signaling can be costly, Study 2 tests whether individuals strategically avoid interactions that could necessitate this type of signaling response. Players play two rounds of dictator games of increasing size, swapping roles in between. In treatments that allow it, many players reject allocations from their partner in the first round; they then act more selfishly as the dictator in the subsequent, higher-stakes round. Together, these results emphasize that the need to signal reciprocity or other socially desired behavior can influence how people engage with and respond to others in strategic contexts.

Keywords: Reciprocity, Fairness, Obligation, Social Image, Saving Face

JEL: C70, D91, M31

* 24 Hopkins Hall Drive, Schapiro Hall, Williamstown MA 01267. mc20@williams.edu. Corresponding author.
^ Division of Social Science, Saadiyat Island, Abu Dhabi, UAE. jchapman@nyu.edu.

I. Motivation and Literature

Social image concerns are a key motive for complying with and signaling socially desirable behaviors (DellaVigna et al. 2012). When these image motives are not present, such as when actions in trust and dictator games are unobserved and there is plausible deniability over who caused a selfish outcome, many individuals are willing to deviate from norms of fairness and reciprocity (Andreoni and Bernheim 2009; Tadelis 2011; Dana et al. 2007). However, other individuals may nevertheless prefer that their actions be observed in order to publicly signal that they acted fairly or reciprocally. These types may view hidden actions not as an opportunity to act selfishly without repercussion, but rather as an obstacle to signaling preferences for fairness, reciprocity, or other socially desired attributes. Consequently, when actions are hidden, they may be willing to pay to signal these preferences to others. Similarly, they may desire to avoid interactions that induce the need to send these costly signals in the first place. This paper extends the social preferences literature by testing for these motives across two experiments.

Many cultures emphasize the importance of publicly signaling socially expected or desired behaviors, such as reciprocity to others. Social customs in Chinese cultures emphasize that gifts must be publicly reciprocated (Steidlmeier 1998); travel websites advise that when in Chinese cultures, you must always "repay [a] gift with something of equal value" and that if you give a gift first, "don't be surprised if your gift is immediately reciprocated with a gift of equal value" (Mack 2019). When traveling to Japan for business, Payne (2016) advises "bring[ing] a range of gifts for your trip so if you are presented with a gift you will be able to reciprocate." These motives are captured in the concept of "saving face" that originates from Chinese culture, which includes publicly signaling compliance with established norms of reciprocity (Rodgers 2019). As a result, reciprocal gifts or favors may be considered "a sign of not just respect [but…] also obligation" (Chang 2016). Finally, these motives are present in most cultures to varying degrees, including western cultures (e.g. Gergen et al. 1975) and even "archaic" or primitive societies (e.g. Mauss 1925).

These examples demonstrate that reciprocity can be motivated, at least partially, by the need to signal that one has reciprocated. This motive is consistent with economics models suggesting that actions

can signal beliefs or types (e.g. Levine 1998; Gul and Pesendorfer 2005; Benabou and Tirole 2006; Charness and Dufwenberg 2006; Sliwka 2007). For instance, reciprocating to gifts in these cultural contexts is a signal that you are the type who abides by that culture's norms, or that your beliefs and preferences align with these norms. Games with hidden actions, as in Andreoni and Bernheim (2009) or Tadelis (2011), may interfere with the ability to send this signal, to the dismay of some.

Study 1 in this paper tests whether some people are willing to pay to signal reciprocity in a game with hidden actions. Players play a modified trust game where the second-mover's allocation can be stochastically prevented by nature. Specifically, the first-mover makes a binary choice over whether to keep or give $1, with a standard 3x multiplier for giving. The second-mover chooses, via strategy method, what portion of the $3 to give back; however, the second-mover's choice is overridden by nature with 50% probability. When nature intervenes, it allocates $0 back, but the first-mover does not know whether it was their partner or nature that was responsible for their $0 outcome.

After making their allocation, second-movers are given the option (via strategy method) to pay $0.10 to truthfully inform their partner what they attempted to allocate. This allows them to signal their intent to reciprocate in the event that nature intervenes and allocates $0. Crucially, the ability to send a message is *not* common knowledge, and the second-mover is aware that declining to send a message does not convey any information to the first-mover. Using the strategy method, we find that 44% of second-movers were willing to pay $0.10 to send this message if they received the $3 from the first-mover.

Importantly, message sending was much lower in a baseline condition where the first-mover's choice was replaced with a (virtual) coin flip. In this condition, the second-mover received the $3 at random (i.e. chosen by nature) instead of due to the first-mover's choice, and this fact was common knowledge to both players. Rates of message sending dropped to only 17%, even though the only difference was whether receiving the $3 was attributed to nature or to the first-mover. Therefore, knowledge of the first-mover's intention to give led to greater willingness to send a message. This indicates that many players sent messages to signal that they tried to reciprocate to their partner's intentions.

This desire to signal reciprocity, or other similar socially desired behavior, can complicate decisions in social settings. For instance, there can be significant anxiety or social pressure to reciprocate and repay an obligation (Xiong et al. 2018). Articles describing gift-giving practices in China refer to receiving a gift as simultaneously a "blessing and a curse" (Chang 2016) because of the resulting expectations of a reciprocal action. These motives can be burdensome in strategic contexts, such as when making business decisions that could impact others that you desire to reciprocate to. As such, saving face may involve not only promptly and publicly reciprocating when necessary, as seen in Study 1, but also avoiding situations that could necessitate such a response in the first place.[1]

Study 2 tests whether people will avoid actions or outcomes that would otherwise raise expectations of an action or signal in response. Players play two sequential rounds of dictator games of increasing size ($0.50 and then $2), with partners swapping roles in between rounds. In one condition, recipients in the first round are allowed to reject (and return) the dictator's choice to give the $0.50. If rejecting the $0.50 enables them to avoid feeling obligated to act equitably or reciprocally in response, then it may enable them to act more selfishly as the dictator in the subsequent larger-stakes round. We find results consistent with this mechanism; 20% of subjects choose to reject when allowed, and offering the option to reject leads to more selfish actions in the subsequent $2 dictator game and higher total pay for the second round dictator.

The results from both studies add to our understanding of social image and preference signaling. Similar to previous literature (e.g. Dana et al. 2007; Andreoni and Bernheim 2009; Tadelis 2011), Study 1 finds that some subjects will take advantage of plausible deniability to act selfishly; however, we also identify many subjects who still reciprocate positively despite plausible deniability, and many of them choose to pay to signal that they did so.[2] Therefore, there exists a type who does not take advantage of plausible deniability to act selfishly, instead acts reciprocally, and is also willing to pay to ensure that this

[1] Such behavior is also noted by Jamaican songwriter Bob Marley: "Never make a politician grant you a favor; they will always want to control you forever" (*Revolution*). Many thanks to an anonymous referee for this suggestion.
[2] Dana et al. (2007), Andreoni and Bernheim (2009), and Tadelis (2011) also all found some subjects who did not act selfishly despite the hidden actions, although none of these studies directly measured whether those individuals were willing to pay to signal that they did not act selfishly.

is signaled to others. Study 2 builds on this by demonstrating that some individuals will, when allowed, choose to reject pay in order to avoid outcomes that may compel them to act more equitably or reciprocally in response. By rejecting, they are able to act more selfishly in subsequent strategic interactions, leading to higher net pay for themselves.

The results also add nuance to existing models of social preferences. Some models posit that individuals are motivated by a desire to meet others' expectations and thus avoid guilt (e.g. Charness and Dufwenberg 2006; Battigali and Dufwenberg 2007; Blanco et al. 2010). Study 1 suggests that when actions can be obscured, individuals may also be concerned with communicating that they were not the reason that expectations were not met. Doing so may be key to alleviating guilt (or perhaps shame, as suggested by Tadelis (2011)), although it should be noted that our results do not directly measure for guilt or shame.

II. Study 1 – Costly Signaling

2.1 Experimental Design

2.1.1 Trust Game Overview

Our first experiment uses a modified trust game to test whether individuals are willing to pay to signal reciprocity. We adopt a similar game as Tadelis (2011), where actions of the second-mover in the trust game are overridden by nature with some probability. We add a novel design element by allowing the second-mover to send (via strategy method) a costly message to their partner that reveals what they attempted to allocate, in the event that nature overrides their choice. This gives the second-mover a chance to signal their intention to reciprocate even when actions are hidden. This message option is not known to the first-mover, and the second-mover is aware that the absence of a message conveys no information to their partner.

2.1.2 Trust Game Conditions

Participants were randomly allocated to one of three conditions. Each condition corresponds to different versions of the trust game, as follows.

We designate the main treatment condition as the *player-choice: cost-self* condition. In this version, the first-mover (labeled Player A in this section) is given an endowment of $1, and they can choose to keep or give the $1 to their partner (Player B). If given, the $1 triples to $3. Player B is asked, via strategy method, how much of the $3 they would give back to Player A if they receive the money. Player B can enter any amount between $0 and $3.

In contrast to the standard trust game, nature intervenes and overrides Player B's choice with 50% probability. This is explained to subjects as a 'virtual coin flip' made by the computer (see Figure 1). If nature intervenes, then $0 is returned to Player A and Player B keeps all of the $3. Since Player A is not told whether nature intervened, they do not know whether a $0 outcome is because of their partner's choice or because nature intervened (and this is common knowledge). As in previous literature (Tadelis 2011), this design gives Player B the opportunity to act selfishly and give $0 without appearing selfish to others. However, it also serves as an obstacle to those that instead want to signal that they want to reciprocate.

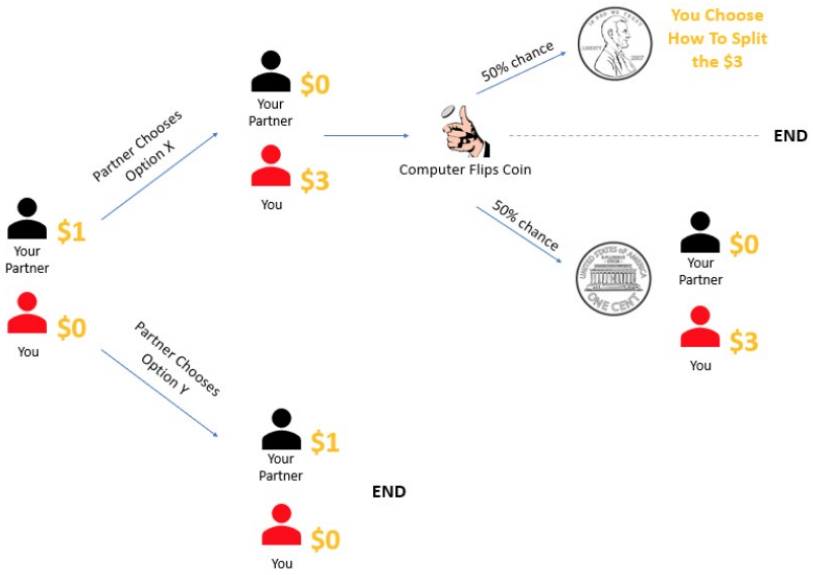

Fig 1. Graphic shown to Player B as part of the instructions (*player-choice: cost-self* condition). "Option X" is to give the money, while "Option Y" is to keep it.

Player B then chooses (via strategy method) whether to pay a fixed cost of $0.10 to inform their partner what they attempted to allocate, in the event that nature intervenes (see Figure 2). This message provides Player B with a chance to convey that they tried to send money to their partner, but that nature

prevented it. They are informed that the message will be sent (and they will pay the $0.10) *only* if their partner actually gives the $1 and nature actually intervenes to override their allocation. Importantly, Player A is not aware of the option to send a message, and thus they will not infer any information from the absence of a message; this is made clear to Player B in their instructions. Finally, Player B is not made aware of the message option until after they have made their allocation of the $3; the message option therefore cannot influence allocations.

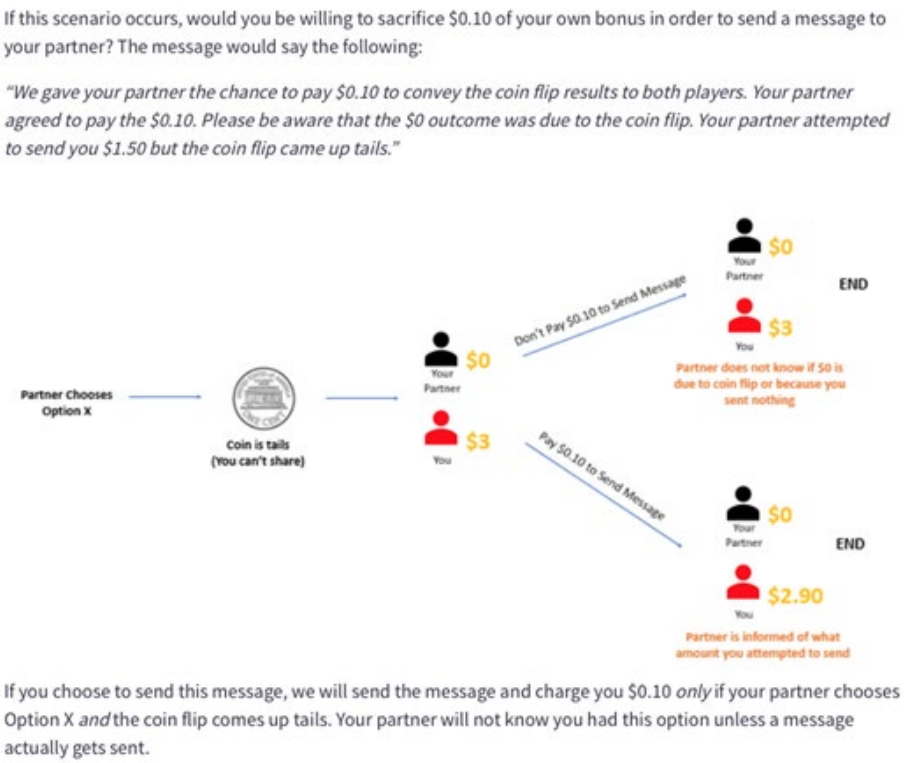

Fig 2. Graphic shown to Player B instructing them about the message (player-choice: *cost-self* condition); in this example, Player B had previously chosen to allocate $1.50.

A second treatment condition, termed the *player-choice: cost-both* condition, changes the cost of the message. All aspects are identical to the *cost-self* condition except now the message costs *both* players $0.10 apiece. This treatment tests whether Player B is still willing to send the message even when it also deducts pay from Player A, the person they would be signaling to. See Figure 3 for the exact instructions given to Player B, and the exact wording of the message that gets sent.

If this scenario occurs, would you be willing to have both you and your partner sacrifice $0.10 each in order to send a message to your partner? This message would say the following:

"We gave your partner the chance to decide whether both players should pay $0.10 to convey the coin flip results to both players. Your partner agreed to do this. Please be aware that the $0 outcome was due to the coin flip. Your partner attempted to send you $1.50 but the coin flip came up tails."

This choice is represented graphically as:

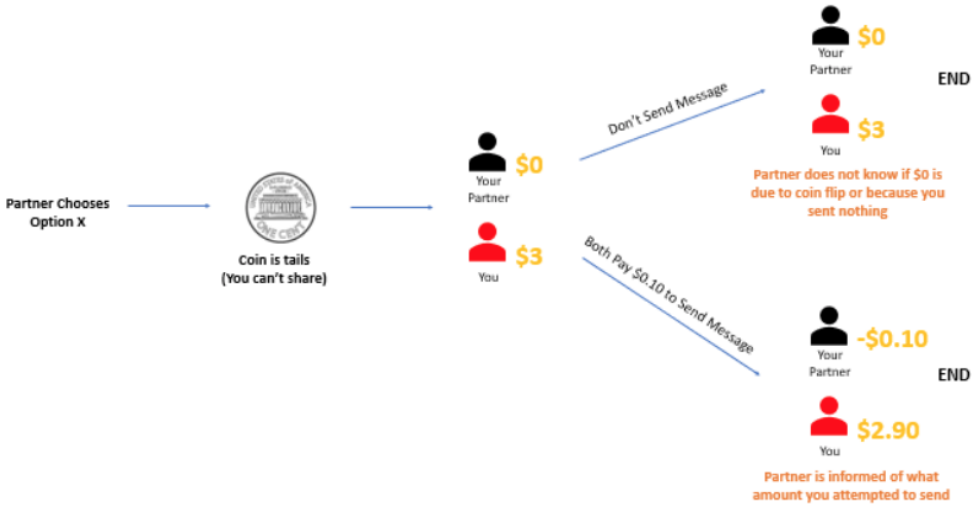

Fig 3. Graphic shown to Player B instructing them about the message (*player-choice: cost-both* condition); in this example, Player B had previously chosen to allocate $1.50.

The third condition, which we term the *nature-choice* condition, replaces Player A's choice with a virtual coin flip. This coin flip (50% chance) decides whether the initial money is given to Player B. Player B then makes the same strategy method decisions as before, including whether to pay $0.10 to send a message indicating what they attempted to allocate. Therefore, the only difference between *nature-choice* and *player-choice: cost-self* is whether receiving the $3 allocation conveys any information about Player A's intentions or preferences. By comparing these two conditions, we can thus isolate whether perceived Player A intentions are a key motive for message sending. (This *nature-choice* counterfactual follows the design from previous reciprocity experiments that isolated intentions from outcomes, e.g. Charness and Rabin 2002; Cox 2004; Falk et al. 2008; Chao 2018).

In all three conditions, all players had to pass a multiple-choice comprehension quiz (customized for each condition) to ensure they understood the details of the game prior to starting the experiment.

2.1.3 Questionnaire

The trust game is followed by a questionnaire. The questionnaire asks subjects to explain (in open-ended responses) the motives behind their choices in the game, as well as demographics questions and a psychological personality scale on anxiety to reciprocate (Xiong et al 2018). It also asks subjects to guess what actions their partner expected them to make (Player B received \$0.10 if they guessed within \$0.05 of their partner's elicited expectations).[3] Subjects received a bonus of \$0.25 for completing the questionnaire.[4] Despite being only correlational and self-reported, we collected these variables in case they provided additional insights on motives.

2.1.4 Experimental Procedure

Subjects were recruited via Amazon's Mechanical Turk online platform in March 2019. Subjects were required to have a U.S. IP address and to have completed 100+ previous Human Intelligence Tasks ("HITs") with a 98% or better acceptance rate. Subjects were paid a base pay of \$0.50, plus a \$0.25 bonus for filling out the questionnaire, and an additional bonus between \$0 and \$3 depending on the outcome of the game. A total of 294 subjects completed the experiment across the three conditions.

Subjects were paired and completed the experiment on the iDecisionGames online platform. The Mechanical Turk HIT linked directly to the iDecisionGames exercise, which then randomized subjects into the three conditions. Once paired, subjects were able to answer all questions up until the pay screen, even if their partner was slow or dropped the HIT (since the game was done entirely via strategy method).[5]

[3] Since expectations are elicited after allocations, reverse causality is a possibility; allocations could have influenced reported expectations, instead of the reverse. Nevertheless, this was preferred to eliciting expectations before allocations, since we did not want to influence allocation decisions via the elicitation.

[4] Since all Player As earned this \$0.25 bonus, no Player A ever received a net negative bonus, even if their partner chose to send the message in the cost-both condition.

[5] The player finishing first was shown a wait-screen that included the mTurk HIT code. They could choose to wait for their partner to see their bonus (and the message, if sent); those that did not were informed of their pay (and the message, if sent) when bonuses were paid. In rare cases where a partner did not complete the experiment, another player's choices were randomly selected to determine pay. Since a player could finish the experiment even if their partner dropped, 146 Player Bs and 148 Player As completed the experiment. Incomplete responses were discarded without being analyzed.

Figure 4 depicts how experimental sessions proceeded. Subjects signing up for the HIT were randomly assigned to one of the three conditions, and then to a role, before completing the game and questionnaire. Full subject instructions and screenshots are available in the Supplemental Materials.

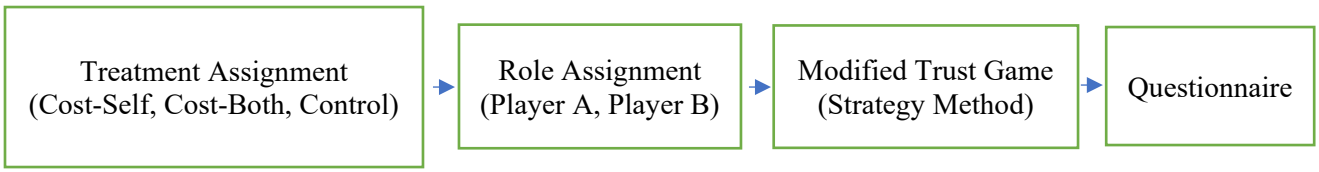

Fig 4. Study 1 procedure.

2.2 Results and Interpretation

2.2.1 Main Results

A significant proportion of participants were willing to send the message, as shown in Figure 5. Eight of 48 subjects (17%) chose to pay to send the message in the *nature-choice* condition (greater than zero, one-tailed t-test, p = 0.002).[6] A much higher proportion, 21 of 48 subjects (44%), sent messages in the *player-choice: cost-self* condition. The difference between the two conditions is statistically significant (two-tailed, p=0.004). In the *player-choice: cost-both* condition, 12 out of 50 (24%) subjects were willing to send the message (greater than zero, p = 0.000). This latter result indicates that subjects sent messages even when doing so reduced pay for not just the sender but also their partner that they were signaling to.

In all conditions, average allocations were higher among message-senders than non-senders. Senders (n=41) allocated $1.23 on average while non-senders (n=105) allocated $0.71 on average (difference-in-means, p=0.000). Amongst message-senders, average allocations did not differ between *nature-choice* and *player-choice* ($1.31 vs. $1.20, n=41, p=0.590). Among non-senders, allocations were significantly lower in *nature-choice* ($0.46 vs. $0.86, n=105, p=0.003). Altogether, average allocations were therefore higher in *player-choice* conditions ($0.98 vs. $0.60, p=0.002) due to having a higher proportion of message senders, as well as higher allocations among non-senders.

[6] We use one-tailed tests when comparing proportions against zero, since proportions (i.e. rates of message sending) cannot fall below zero. These are always labeled "greater than zero" before the listed p-value. All t-tests in this paper not labeled in this way are two-tailed.

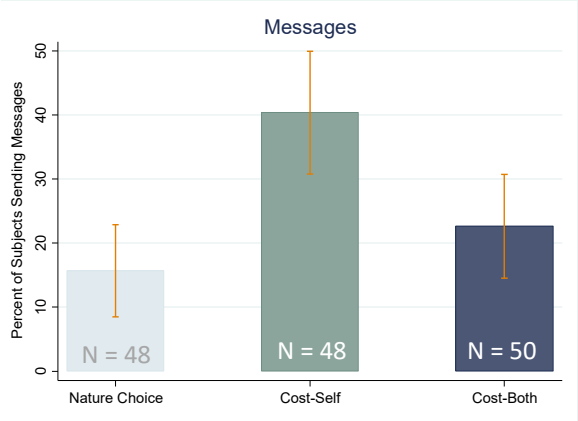

Fig 5. Message sending in nature-choice and cost-self conditions.
Bars represent 95% CI.

2.2.2 Interpretation and Mechanisms

Our main result relies on the differences in Player B choices between the *cost-self* and *nature-choice* conditions. Specifically, Player Bs allocated more, and then were more willing to send the message, in the *cost-self* condition relative to the *nature-choice* condition. Since the only difference between these conditions is whether the initial money transfer was attributed to Player A or to nature, the increased allocations and message sending can be attributed to differences in perceived Player A intentions. In other words, when Player B knows they received the $3 due to Player A's intent instead of due to nature's choice, they allocate more, indicating intentions-based reciprocity; they are then more willing to pay $0.10 to send the message, reflecting the message's ability to serve as a signal of reciprocity to Player A's intentions.[7]

In the *cost-both* condition, twelve out of 50 (24%) Player Bs were willing to send the message. This rate is greater than in the *nature-choice* condition, but not significantly so (p=0.373). This is possibly because, while the *player-choice* aspect of the condition increased desire to send the message (as seen in the *cost-self* condition), the higher cost of the message relative to *nature-choice* likely counteracted this. Nevertheless, many were willing to send the message, even when it cost their partner $0.10 and even when

[7] An alternative interpretation could be that any time someone allocates a high amount, they choose to send the message (e.g. perhaps to signal altruism), and the *player-choice* condition has higher message sending simply because more subjects allocate higher amounts. However, since the *player-choice* condition has higher allocations only because they reciprocated to Player A intentions, it is more likely that the increase in message sending is due to a desire to signal this reciprocity to intentions; after all, reciprocity to Player A intent is the reason allocations increased in the first place.

their partner just earned $0 in the trust game. This indicates that for some, the desire to send the message superseded the desire to maximize pay for the very person they were signaling to.

The post-experiment questionnaire suggests that the choice to send a message was not driven by differences in beliefs about their partner's expectations. In the questionnaire, Player Bs were incentivized to guess what their partner would expect them to give back in the second stage, assuming they were given the money in the first stage. Average guesses for Player Bs were not different between message-senders and non-senders (both averaged $1.26, p=0.976).[8] Differences were also insignificant when comparing senders and non-senders within just *nature-choice* ($1.03 vs. $1.01, p=0.950) or just *player-choice* ($1.41 vs $1.32, p = 0.329) conditions. Since message-senders did not differ from non-senders along this dimension, it could instead be that they placed greater weight on what they thought their partner expected, or that they simply cared more about their image and thus placed greater importance on signaling their preference or type.

Since eight subjects (17%) sent messages in the *nature-choice* condition, some were willing to pay for messages even when Player A intentions were not clear. Message-senders in this condition may have wished to signal fairness or equity to their partner, since they allocated $1.31 on average (compared to $0.46 for non-senders in the same condition). In fact, more than half of the senders (5 out of 8) allocated an exact even split of $1.50, which is consistent with acting according to norms of equality. Alternatively, they could be reciprocating to their partner's change in outcome (either without considering intentions, or assuming they may have intended to give), and then signaling via the message that they attempted to reciprocate. Finally, it could also be as simple as a signal of altruism. Although multiple motives for message sending in this condition are therefore plausible, this does not interfere with our ability to conclude that differences in message sending between the *nature-choice* and *player-choice: cost-self* conditions are nevertheless attributable to differences in perceived Player A intentions, and thus to intentions-based reciprocity.

[8] These high average expectations are also consistent with expectations of reciprocity in general; Player B's on average believed their partners expected more than the initial $1 to be given back.

Finally, our results do not directly speak to whether other mechanisms can also motivate message sending in other contexts. For instance, as possibly suggested by message sending in the *nature-choice* condition, signaling fairness or equality may also be a motive, although our results do not isolate this mechanism with certainty. Likewise, changes in outcomes (such as increasing the 3x multiplier in the trust game) or in the cost of the message could also change willingness to engage in costly signaling. Study 1 was only designed to detect whether intentions-based reciprocity can motivate messages and signaling, although these other mechanisms are worthwhile questions for future research.

III. Study 2 – Avoiding Obligation

3.1 Experimental Design

3.1.1 Motivation and Game Overview

Study 1 suggests that individuals often wish to signal reciprocity after receiving a gift or kind action. The desire to do so may be reflective of obligation, social pressure, or anxiety to reciprocate (Xiong et al. 2018). Rather than be subjected to these concerns, some individuals may be motivated to avoid actions or outcomes that necessitate such a response. This may be especially true in strategic contexts, such as business or political settings where gifts can be used to influence subsequent decision-making.

Study 2 tests for this rejection behavior by having players participate in two sequential rounds of dictator games with increasing stakes. In the first round, the dictator makes a binary choice of whether to give or keep $0.50. In one condition, the receiver in this first round is allowed to reject what the dictator gives them; if they do, the amount is returned to the dictator. Regardless of whether they reject, the receiver then switches to become the dictator in a second round with larger stakes of $2, and they choose how much of that $2 to share with their partner. If rejecting allows a player to avoid feeling obligated to act fairly or reciprocally, they may choose to reject in the first round in order to act more selfishly in the subsequent, higher-stakes round. Thus, relative to conditions where rejecting is not possible, being allowed to reject may raise Player B's total pay across the two rounds of the game.

3.1.2 Dictator Game Conditions

In our *baseline: player-choice* condition, the dictator in the first round, designated Player A in this section, is given an endowment of $0.50 and a binary choice to keep or give the $0.50 to their partner, Player B.[9] Since this is a baseline condition, Player B is **not** allowed to reject the $0.50 if given. In the second round, Player B is the dictator, the pie size is $2, and they can give any portion of that $2 to Player A. Player B makes their choice using the strategy method; first they choose what they would allocate if they were given the $0.50, and then they choose what they would allocate if they were not given the $0.50. All of these details are common knowledge to both players. As in Study 1, instructions are presented in part using a graphic (see Figure 6).

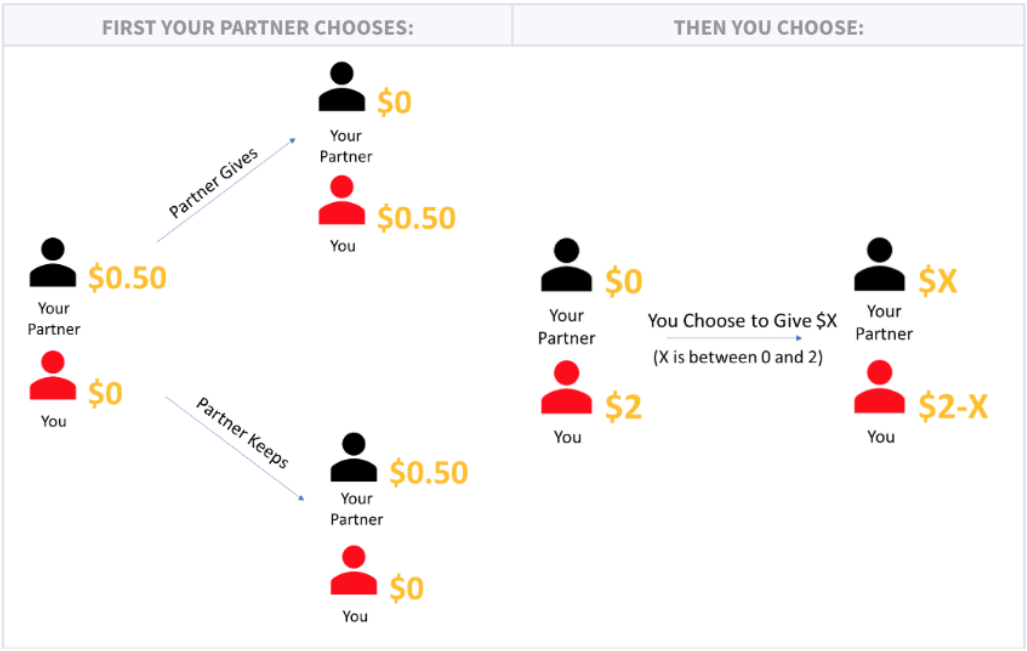

Fig 6. Graphic instructions to Player B, *baseline player-choice* condition.

In the *reject: player-choice* condition, Player B is allowed to reject (via strategy method) the initial $0.50 gift from Player A, if it is given (see Figure 7). If accepting the $0.50 from their partner compels or obligates them to act more fairly or reciprocally in the second round, they may choose to reject the $0.50 in order to subsequently act more selfishly, leading to higher net pay for Player B across both rounds.

[9] Player A's choice is binary in order to allow strategy method elicitation of Player B's second round allocation.

Therefore, this mechanism would predict that Player Bs reject at rates significantly greater than 0%, and subsequently earn more total pay across both rounds than when they are not allowed to reject.

The remaining two conditions implement *nature-choice* baseline conditions, similar to the one seen in Study 1. In the *baseline: nature-choice* condition, Player A no longer decides whether to give or keep the initial $0.50 in the first round. Instead, nature makes the choice and chooses whether to give with 50% probability. As in Study 1, this helps to separate whether Player A's intentions have any influence on allocations. The fourth condition, the *reject: nature-choice* condition, combines both the reject and nature-choice manipulations simultaneously to test whether Player A's intentions matter for the choice to reject.

As in Study 1, subjects had to pass a multiple-choice comprehension quiz (customized to each condition) about these instructions to proceed to the game.

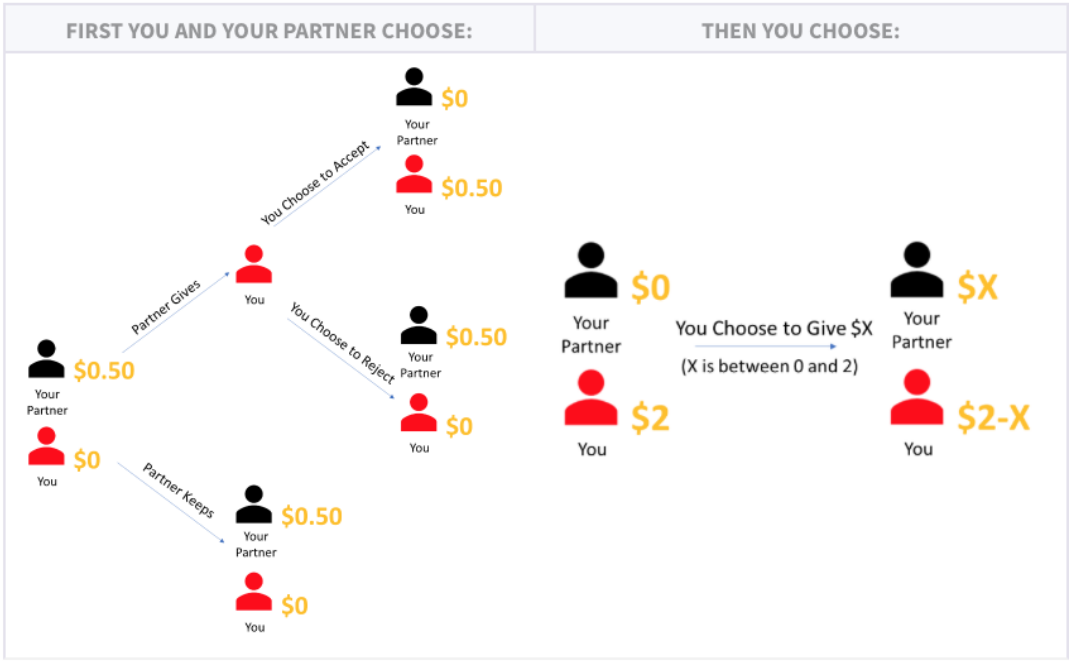

Fig 7. Graphic instructions to Player B, *reject player-choice* condition.

We switched from the modified trust game in Study 1 to this sequential dictator game in Study 2 in order to enable the *reject* option. In a trust game, rejecting would mean that Player B no longer has any money to allocate back in the subsequent round. To measure how players allocate after rejecting, we therefore needed an alternative design that endowed Player B with money to allocate even when choosing to reject in the first stage.

3.1.3 Questionnaire

As in Study 1, the game is followed by a questionnaire asking subjects (in open-ended responses) to identify subject motives and expectations during the game, as well as demographics questions and the same psychological personality scale. As before, these are correlational and self-reported variables only; they are not part of the main analysis but may provide some additional insights into motives.

3.1.4 Experimental Procedure

Subjects were recruited via Amazon's Mechanical Turk online platform in July 2019 using the same criteria as in Study 1. Subjects that participated in Study 1 were excluded. Subjects were paid a base pay of $0.50 with a chance to earn an additional bonus between $0 and $2.50, depending on the outcome of the dictator games. A total of 388 subjects completed the experiment, with 196 assigned to Player B.[10]

The experiment was programmed in iDecisionGames. As in Study 1, the Mechanical Turk HIT linked to the program, which randomized subjects into conditions and allowed real-time pairings. Once paired, subjects were able to answer all questions up until the pay screen, regardless of how fast their partner was or whether their partner dropped the HIT.

Figure 8 depicts how experimental sessions proceeded. Subjects were assigned to one of the four conditions, and then assigned as either Player A or Player B before completing the game and questionnaire. Full instructions are available in the Supplemental Materials.

Fig 8. Study 2 procedure.

[10] As in Study 1, since it is strategy method, players could proceed to the end of the study without waiting for their partner. In the rare cases where one's partner dropped the HIT, pay was again determined by using the choices of a randomly chosen participant.

3.2 Results and Interpretation

3.2.1 Main Effects

Players rejected allocations in both conditions that allowed it, as seen in Figure 9. In the *reject: nature-choice* condition, 15 out of 52 (29%) subjects turned down a payment of $0.50 from the dictator (different from zero, p=0.000), while in the *reject: player-choice* condition, 9 out of 46 (20%) did so (different from zero, p=0.002). Rejection rates did not differ between *nature-choice* and *player-choice* (two-tailed, p=0.291); we discuss possible reasons for this in the subsequent interpretations section.

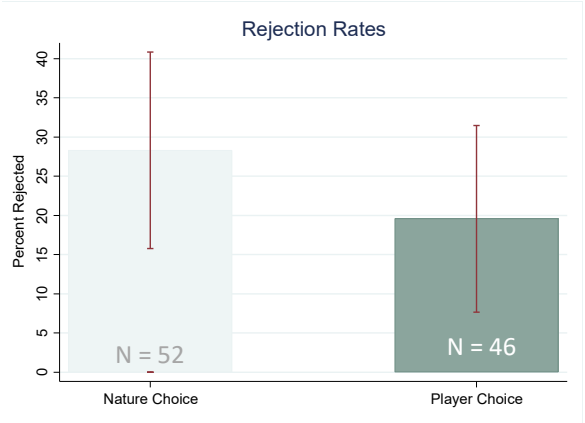

Fig 9. Rejection rates, *nature-choice* versus *player-choice.*

There is evidence that the ability to reject increased average total pay for Player Bs. We test this using OLS regressions that combine data from all conditions and analyze Player B allocations when offered the $0.50. The main explanatory variable is an indicator for whether rejections were allowed. The main dependent variable is either how much Player B gave to their partner in the second round (Model 1), or how much total pay Player B earned across both rounds (Model 2). The regression specification is thus:

$$R_i = \beta_0 + \beta_1 * player\text{-}choice_i + \beta_2 * reject\text{-}allowed_i + \varepsilon_i$$

In this specification, *i* indexes Player Bs, *R* represents Player B's allocation, *player-choice* is an indicator that equals 1 if Player A was responsible for the first-round choice, and *reject-allowed* is an indicator that equals 1 if Player B was given the option (via strategy method) to reject the first allocation.

Model (1) demonstrates that having the reject option significantly reduced Player B allocations to Player A when they were initially offered the $0.50. However, in principle this could merely reflect Player

Bs rejecting the $0.50 and then allocating $0.50 less in the second round, leaving total allocations unchanged. The results of Model (2), where the dependent variable is Player B's total pay, demonstrate that this was not entirely the case. On average, Player Bs earned $0.11 more across both rounds when given the option to reject. Importantly, these are intend-to-treat effects, since we are estimating the effect of being allowed to reject, and not conditioning on whether rejections took place. Since only 20% of subjects rejected in the *player-choice* condition (and 29% in *nature-choice*), treatment effects on those actually choosing to reject are likely larger.

Table 1: OLS Regressions

	(1)	**(2)**
Dependent Variable	**Second Round Allocation from Player B to Player A**	**Player B Total Pay**
Observations	**196**	**196**
Player B Offered $0.50	**Yes**	**Yes**
R^2	**0.085**	**0.025**
Player Choice	0.115*	0.092
	(0.063)	(0.067)
Reject Allowed	-0.234***	0.113*
	(0.063)	(0.067)
Constant	0.666***	1.731***
	(0.054)	(0.054)

*p<0.10; *** p<0.01. Robust standard errors in parentheses.

Rejections may have affected pay by changing Player B's willingness to split the $2 evenly. In the *baseline: player-choice* condition, 19 of 51 players (37%) split evenly if they were first offered the $0.50, but in the *reject: player-choice* condition only 8 out of 46 (17%) did so (difference-in-means, p = 0.029). In addition, none of the subjects who actually rejected in the *player-choice* condition chose to split evenly. Taken together, it appears some subjects may choose to split the $2 evenly if they receive the $0.50 from their partner, as seen in the *baseline: player-choice* condition, but if given the chance would prefer to reject the $0.50 and allocate significantly less than half instead. In the *nature-choice* conditions, 15 out of 47 (32%) allocated half when given the $0.50 in *baseline*, compared to 11 out of 52 (21%) in *reject*, but this difference does not reach significance (p = 0.229); perhaps this is because the *nature-choice* conditions invoke less reciprocity, since Player A intentions cannot be attributed to the $0.50 transfer.

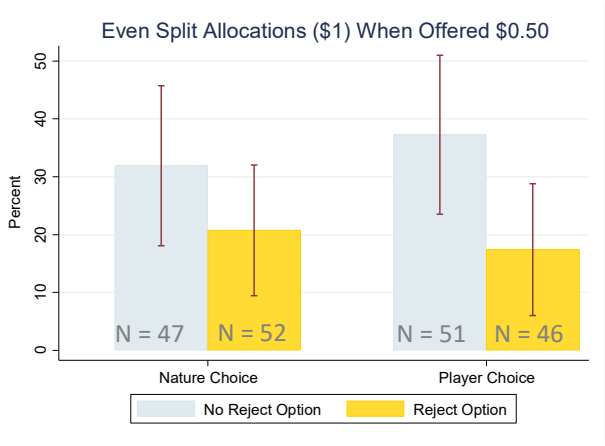

Fig 10. Rates of even splits, by condition.

Finally, very few Player Bs chose to split evenly if they were not first given the $0.50. Across both *player-choice* conditions, only two Player Bs split evenly if Player A did not first offer the $0.50 (compared to 27 otherwise), and only nine did so across the *nature-choice* conditions (compared to 26 otherwise). In short, not being offered the $0.50 also led to fewer even splits of the $2.

3.2.2 Interpretation

The results can be explained by preferences for fairness, preferences for reciprocity, or both. After receiving the $0.50, Player Bs may be more inclined to act fairly and implement equal distributions. The data demonstrates that when not offered the $0.50, or when the $0.50 could be rejected, there were significantly fewer subsequent equal splits of the $2. Thus, first receiving the $0.50 appears to be key for generating equal distributions of the $2, and rejecting may allow Player Bs to avoid this response. Alternatively, instead of fairness, it could be a form of outcomes-based reciprocity; Player Bs could feel obligated to reciprocate after receiving $0.50 that came out of Player A's bonus, and rejecting enables Player Bs to avoid having any change in outcome to reciprocate to. Since allowing rejections leads to such stark differences in the rate of equal splits, we are inclined to view this as stronger evidence of a fairness and equality mechanism than a reciprocity mechanism, but we cannot definitively distinguish between the

two. Nevertheless, we can still conclude that subjects find reason to reject the $0.50 when allowed to do so, and the ability to reject leads to outcomes that are simultaneously less fair and less reciprocal.[11]

Unlike in Study 1, the results do not suggest that perceived intentions matter for rejections. This is because rejection rates did not differ between the *player-choice* and *nature-choice* versions of the reject conditions, despite the difference in perceived intentions of Player A. One possibility is that the amount, $0.50, is simply too small (unlike in Study 1, where the amount given was effectively $3), and thus the intention to give it does not provide a strong enough signal of Player A's intentions or type. Consistent with this, Model (1) estimates that receiving the $0.50 from Player A instead of from nature only increased allocations by $0.12 on average (p=0.071), implying there was only marginal positive reciprocity in response to Player A's intentions to give. Similarly, some could have viewed giving the $0.50 as a strategic choice or bribe, instead of kindly intended.[12] If so, receiving the $0.50 likely would not invoke much positive intentions-based reciprocity, regardless of whether it was *nature-choice* or *player-choice*.

Another possible reason why rejection rates did not differ could be that subjects felt it would be "rude" to reject a payment given by another player. This would not be as likely to apply when the $0.50 is instead given as a result of nature's choice. Thus, perhaps an increased desired to avoid accepting the $0.50 in the *player-choice* condition was counterbalanced by increased social pressure to accept. Consistent with this mechanism, five subjects in the *reject: player-choice* condition stated in the questionnaire that they did not reject because they felt it would be rude to do so.[13] In contrast, no subjects in the *reject: nature-choice* condition expressed this sentiment.

---

[11] If rejecting does not enable Player Bs to be less fair or reciprocal, there may not be much reason to reject. For instance, they can accept the $0.50 and then simply return it by allocating $0.50 more in the second round. This would only be infeasible if Player B initially intended to give more than $1.50 in the second round (thus making them unable to increase their allocation by $0.50); however, no Player B ever gave more than $1.50 after rejecting, suggesting this was never a motive for rejecting.

[12] They may be less likely to assume the gift was a bribe in Study 1, since the 3x multiplier in the trust game could make giving the initial $1 seem partially driven by preferences for maximizing total welfare. Giving $1 in Study 1 could also be perceived as altruism or fairness since (unlike in Study 2) Player B receives nothing unless the first-mover chooses to give.

[13] Responses were coded by a research assistant who analyzed the responses without regard to condition. The RA was tasked only with categorizing responses, and she was given discretion to choose the categories based on the responses she saw. She placed five responses into a "rude to reject" category; none were from the nature-choice condition. The responses were, in no particular order: (1) At first I was going to reject in order to make the math easier. But then I

IV. Discussion

In experiments where nature obscures player actions, existing literature emphasizes the existence of types that use this plausible deniability to act more selfishly (Andreoni and Bernheim 2009; Tadelis 2011). Study 1 finds evidence of another type that views this plausible deniability not as an opportunity, but rather as an obstacle to their desire to signal reciprocity. These individuals still allocate positive amounts in a trust game despite the ability to "hide" their actions, and they subsequently demonstrate a willingness to pay to signal their reciprocity. The results find that perceived intentions matter to this costly signaling; message sending rates were much higher when their partner and not nature was responsible for giving them their initial money. Finally, some individuals were willing to cost both themselves *and* the person they were indebted to in order to send this signal, suggesting that signaling superseded maximizing pay for their partner.

Study 2 extends these results by demonstrating that some individuals will reject money in order to avoid situations that might necessitate costly signals or actions. Giving individuals the option to reject a gift leads to more selfish and less equitable choices in subsequent interactions. Importantly, Study 2 results are not as robust as those in Study 1, since the design engendered only small levels of positive reciprocity in subjects. In addition, the results cannot definitively isolate whether norms of fairness or reciprocity were more relevant to subjects' reasons for rejecting. Careful experimental design in follow-up research can improve upon these results.

The behaviors identified in this paper help extend our understanding of social preferences to the concept of "saving face." The behavior in Study 1 underscores that the importance of signaling reciprocity can at times outweigh the desire to maximize our own pay or even the pay of the one we are signaling to. This latter effect is consistent with gift-giving customs where individuals stash many gifts in advance in order to reciprocate at a moment's notice, even though these gifts are not personalized or catered

thought 'sometimes' people find it rude if you reject their "gift." Which is way overthinking for this scenario. (2) I feel like rejecting it is rude. (3) I wanted to accept it. I thought it would be rude to reject it. (4) well if offered I would definetly accepy. [sic] Common curtiousy [sic]. (5) I figured I'd accept it, as it almost seems rude to give it back. (Responses for other subjects are available in the data set).

specifically to the preferences of its recipient (Payne 2016). Similarly, the behavior in Study 2 is consistent with cultures where gift-giving can be seen as not just a blessing but also a "curse" because of the resulting need to publicly respond (Chang 2016). As a result, individuals may be motivated to avoid outcomes that obligate them to respond, thus enabling them to act more freely in subsequent interactions.

These results may also provide insights on policies that manage conflicts of interest involving reciprocity. For instance, in the United States, evidence demonstrates that doctor prescriptions are influenced by gifts and meals from pharmaceutical sales representatives (Larkin et al. 2017; Chao and Larkin 2019), and sales representatives explain that this is partly because they have access to bi-weekly, doctor-level prescriptions that they can use to "guilt" doctors into reciprocating (Fugh-Berman and Ahari 2007). Study 1's results suggest that prohibiting salesperson access to doctor-level prescriptions data (as many countries already do) could alter this dynamic, since it would remove the signal value of those prescriptions. Similarly, Study 2's results also carry implications for conflicts of interest. For instance, in South Korea, the anti-corruption "Kim Young-Ran Act" of 2016 prohibits public officials from accepting gifts of more than 100,000 won (90 USD) and dinners of more than 30,000 won (27 USD) in order to avoid being unduly influenced. Although primarily targeted at corruption, it is also relevant for those who are not corrupt and do not want to be influenced; the law provides them with a reason and method for rejecting such gifts (when it otherwise might be considered culturally taboo or rude to do so), thus avoiding the subsequent pressure to respond. The results in this paper may therefore provide some insights on how preference signaling can matter in these policy and conflicts-of-interest contexts.

Funding support. Research funds for subject payments were provided by Williams College and New York University Abu Dhabi.

Author declarations of interest: none.

Acknowledgements. We would like to thank Andreas Lange, Sarah Jacobson, Rebecca Morton, Joseph Wang, two anonymous referees, and audiences at WESSI, NTU, and NEEEW for useful feedback and insight. We thank Ally Isley and Yun Jie Song for excellent research assistance. We also thank Rita Anpilogova, Igor Malakhov, and the staff at iDecisionGames for programming the experiments. All errors are our own.

References

Andreoni, J., Bernheim, BD., 2009. Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects. Econometrica 77, 1607-1636.

Andreoni, J., Rao, J., Trachtman, H., 2017. Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. Journal of Political Economy 125, 625-653.

Battigali, P., Dufwenberg, M., 2007. Guilt in games. American Economic Review 97, 170-176.

Benabou, R., Tirole, J., 2006. Incentives and prosocial behavior. American Economic Review 96, 1652-1678.

Blanco, M., Celen, B., Schotter, A., 2010. On blame-freeness and reciprocity: An experimental study. Universidad del Rosario Faculty of Economics Working Paper No. 85. Available at SSRN: https://ssrn.com/abstract=1703546 or http://dx.doi.org/10.2139/ssrn.1703546

Chang, A., 2016. East vs. West: Gift giving culture. Available at Tutorming.com: http://blog.tutorming.com/expats/gift-giving-culture-china-western. Accessed on August 1, 2019.

Chao, M., 2018. Intentions-based reciprocity to monetary and non-monetary gifts. Games 9, 74.

Chao, M., Larkin, I., 2019. Regulating conflicts of interest through public disclosure: Evidence from a physician payments sunshine act. Working paper.

Charness, G., Dufwenberg, M., 2006. Promises and partnership. Econometrica 74, 1579-1601.

Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. The Quarterly Journal of Economics 117, 817-869.

Cox, J., 2004. How to identify trust and reciprocity. Games and Economic Behavior 46, 260-281.

Dana, J., Weber, R., Kuang, J., 2007. Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. Economic Theory 33, 67-80.

DellaVigna, S., Malmendier, U., List, J., 2012. Testing for altruism and social pressure in charitable giving. The Quarterly Journal of Economics 127, 1-56.

Falk, A., Fehr, E., Fischbacher, U., 2008. Testing theories of fairness – intentions matter. Games and Economic Behavior 62, 287-303.

Fugh-Berman, A., Ahari, S., 2007. Following the script: How drug reps make friends and influence doctors. PLOS Medicine 4(4): e150.

Gergen, K., Ellsworth, P., Maslach, C., Selpel, M., 1975. Obligation, donor resources, and reactions to aid in three cultures. Journal of Personality & Social Psychology 31, 390-400.

Gul, F., Pesendorfer, W. 2005. The canonical type space for interdependent preferences. Mimeo, Princeton University, Department of Economics.

Larkin, I., Ang, D., Steinhart, J., Chao, M., Patterson, M., Sah, S., Wu, T., Schoenbaum, M., Hutchins, D., Brennan, T., Loewenstein, G., 2017. Association between academic medical center pharmaceutical detailing policies and physician prescribing. Journal of the American Medical Association 317, 1785-1795.

Levin, D. 1998. Modeling altruism and spitefulness in experiments. Review of Economic Dynamics 1, 593-622.

Mack, L., 2019. Gift-giving etiquette in Chinese culture. Available at Thoughtco.com: https://www.thoughtco.com/chinese-gift-giving-etiquette-687452. Accessed on August 1, 2019.

Mauss, M., 1925. The Gift: forms and functions of exchange in archaic societies.

Payne, N., 2016. Cross cultural gift giving etiquette. Available at Businessknowhow.com: https://www.businessknowhow.com/growth/ccultural.htm. Accessed on August 1, 2019.

Rodgers, G., 2019. Saving face and losing face. Available at tripsavvy.com: https://www.tripsavvy.com/saving-face-and-losing-face-1458303. Accessed on August 1, 2019.

Sliwka, D., 2007. Trust as a signal of a social norm: Hidden costs of incentive schemes. American Economic Review 97, 999-1012.

Steidlmeier, P., 1998. Gift giving, bribery, and corruption: Ethical management of business relationships in China. Journal of Business Ethics 20, 121-132.

Tadelis, S., 2011. The power of shame and the rationality of trust. Haas school of Business working paper. Available at: http://faculty.haas.berkeley.edu/stadelis/shame_trust_030111.pdf

Xiong, X., Guo, S., Gu, L., Huang, R., Zhou, X., 2018. Reciprocity anxiety: Individual differences in feeling discomfort in reciprocity situations. Journal of Economic Psychology 67, 149-161.