# Bayesian Persuasion and Reciprocity: Theory and Experiment

Pak Hung Au  King King Li*

June 2018

**Abstract**

In a Bayesian persuasion setting (Kamenica and Gentzkow, 2011), a sender persuades a receiver to take an action by designing and committing to disclose information about the receiver's payoff of taking the action. We propose a model that incorporates reciprocity into the Bayesian persuasion setting, using the approach of Falk and Fischbacher (2006). The introduction of reciprocal concerns leads to a number of novel predictions. First, the receiver's response changes continuously in the realized signals. Second, when the prior belief is more favorable, the receiver is more difficult to be persuaded, implying that the sender's optimal persuasion strategy involves more informative disclosure. These predictions are supported by experimental data.

JEL Codes: D72, D82, D83, K40, M31

Keywords: Bayesian Persuasion, Information Design, Reciprocity, Experiment

* Au: Division of Economics, Nanyang Technological University; email: phau@ntu.edu.sg; address: 14, Nanyang Drive, 04-67, Singapore, 637820. Li: College of Business, Department of Economics and Finance, and CityUHK Experimental Economics Laboratory, City University of Hong Kong; email: likingking@gmail.com; address: 9-257, Lau Ming Wai Academic Building, City University of Hong Kong, Hong Kong. We thank Hsiu Hoi Sing and Lam Ho Ming for their excellent research assistance. Financial Support by City University of Hong Kong is gratefully acknowledged.

# 1. Introduction

It has long been recognized in psychology and marketing that reciprocation can play a key role in successful persuasion. In his best-seller "Influence: The Psychology of Persuasion", Cialdini (2006) names reciprocation as the "first universal principle of influence". The idea is that people are found to feel obliged to give back to others the form of a behavior, gift, or service that they have received first. While reciprocation and gift exchange in the form of monetary payoff and effort receive a lot of attention in the economics literature (e.g., Güth, Schmittberger, and Schwarze, 1982; Fehr, Kirchsteiger, and Riedl, 1993; Berg, Dickhaut, McCabe, 1995; Charness, 2004), their roles in the context of persuasion have not been explored. In this paper, we propose a model of Bayesian persuasion that incorporates players' reciprocity, and conduct laboratory experiments to test the model's predictions.

The starting point of our analysis is the model of Bayesian persuasion by Kamenica and Gentzkow (2011). In their setting, a sender tries to persuade a receiver to take a certain action by controlling the information structure through which the receiver can learn about the uncertain payoff of taking an action. Instead of assuming that the players care only about the monetary payoffs, as in the standard Bayesian persuasion setting, we allow the players to have reciprocal preference. In particular, if the receiver perceives that the sender has acted kindly, she would derive utility from returning a favor to the sender. Conversely, if the receiver perceives that the sender has acted unkindly, she would derive utility from punishing the sender. We model the receiver's kindness perception using the approach of Falk and Fischbacher (2006): the perceived kindness depends on the difference in the payoffs between the receiver and the sender. A sender who is willing to act in such a way that gives the receiver a high payoff relative to her own payoff is perceived to be kind.

To illustrate the implications of incorporating reciprocity into the Bayesian persuasion, we consider a simple persuasion problem. The state of nature is binary, say red or black. The receiver chooses between two actions, labelled as R and B, and derives a positive benefit if her action matches the state (i.e., R with red and B with black). The sender derives a positive benefit if and only if the receiver chooses R, so he would try to persuade the receiver to choose red. He designs the information structure through which the receiver can learn about the state. Kamenica and Gentzkow (2011) show that the sender's problem can be formulated as choosing a distribution of posterior beliefs, and can be solved by looking for the concave closure of the sender's payoff function in the realized posterior beliefs. Assuming that the receiver is an expected-payoff maximizer, as in Kamenica and Gentzkow (2011), the problem we consider has the sender's payoff being a step function in the realized posterior beliefs: the receiver chooses R if and only if the posterior belief that the state being red is no less than 0.5.

Consequently, the sender's optimal information structure assigns a positive probability only to posterior beliefs (that the state is red) 0 and 0.5.

If the receiver has a reciprocal preference, the sender's problem is less straightforward. If the sender offers an uninformative posterior 0.5, the receiver's expected payoff is minimized. On the other hand, if the sender offers informative posteriors close to 0 or 1, the receiver has a higher expected payoff. The receiver would naturally consider a sender who offers uninformative posteriors as unkind and a sender who offers informative posteriors as kind. Therefore, a receiver with reciprocal prefernce may be willing to punish the sender by choosing black following posterior exceeding 0.5 only marginally. Moreover, the more the posterior exceeds 0.5, the more willing is such a receiver to choose red to reciprocate the sender's kind behavior, thus implying that the sender's payoff function in realized posteriors can be continuously increasing, instead of being a simple step function.

Perhaps more interesting, by incorporating reciprocity, the receiver's responses and hence the sender's optimal information structure may depend on (ex-post) payoff-irrelevant factors. Specifically, consider an increase in the common prior belief about the state being red. The sender's expected payoff goes up as it is more likely that the realized posterior is favorable for the state being red. Consequently, for any realized posterior, the receiver's perception of the sender's kindness goes down, making them less willing to be persuaded to choose R. Therefore, the sender's payoff function in realized posterior shifts downwards, and the optimal information structure becomes more revealing. This, in turn, allows the receiver to obtain a higher monetary payoff. These findings are in interesting contrast to the standard Bayesian persuasion model without reciprocity, which would predict that the receiver's responses and equilibrium expected payoff are invariant to the prior belief.

We test these predictions by laboratory experiments. To avoid the abstract language (such as conditional distribution function, distributions of posterior beliefs, and Bayes-plausibility) used in the standard Bayesian persuasion setting, we design a laboratory game that is strategically equivalent and is very easy to understand. The design of our experiment is summarized as follows. Players are randomly matched into pairs: player A (sender) and player B (receiver). Play A's objective is to persuade player B to take the action of guessing red. The game has the following steps. First, Player A is given 100 balls in which 100*μ balls are red and 100*(1- μ) balls are black (for some μ between 0 and 1). Player A's task is to allocate the 100 balls into 5 urns (some empty urns are allowed). Then, one urn will be randomly drawn, with the probability being the number of balls in the urn divided by 100. Next, the composition (i.e., number of red balls and number of black balls) of the drawn urn will be announced to player B. In the last stage, one ball will be randomly drawn from the urn, and Player B's task is to guess

the color of the ball drawn. If her guess is correct, player B will receive 40 Hong Kong dollars, and zero otherwise. On the other hand, player A will receive 40 Hong Kong dollars if player B guesses red regardless of the actual color of the ball drawn. In sum, the objective of player A is to design an allocation of balls into urns with the objective of persuading player B to guess red. In Section 3, we explain in greater detail why this ball allocation game is (essentially) strategically equivalent to a Bayesian persuasion game in which the prior of the state is μ.

The equilibrium ball allocation predicted by the standard Bayesian persuasion model is to allocate 100*μ red balls and 100*μ -1 (or 100*μ) black balls in one urn, and the rest (of black balls) into other urns. Since the first urn has about 200* μ balls, there is about 200* μ % chance that it will be drawn. If this urn is drawn, player B will guess red. If the other urn is drawn, player B will guess black.

To test the effect of the prior belief μ on players' behaviors, we run two treatments with μ=0.3 and μ=0.5 respectively. Our main findings are as follows. First, the receiver's response is not a step-function; rather, her probability of guessing red is strictly increasing around the 50% mark. In fact, given the empirical responses of the receivers, the empirically optimal ball allocation differs from the theoretical prediction of Bayesian persuasion described above, while it is consistent with our model in which reciprocity concern is incorporated. Second, for urns with a fraction of red balls around 50%, the receiver's probability of guessing red is strictly lower when μ=0.5 than when μ=0.3. In response to the more demanding receivers, the senders offer urns with more favorable composition when μ=0.5. Our results highlight the importance and empirical relevance of incorporating reciprocity in Bayesian persuasion.

The rest of the paper is organized as follows. A literature review is given below. Section 2 presents our theoretical model and develops our hypotheses. Section 3 contains a detailed description of our experimental design, as well as explaining why the experimental game is strategically equivalent to the Bayesian persuasion game. The main experiment results are reported and analyzed in Section 4. Finally, Section 5 concludes.

## 1.1 Literature Review

Kamenica and Gentzkow (2011) initiated a large and growing theoretical literature on communication game in which the sender can commit to the information acquisition and disclosure policies. Their model has been applied to studying information transmission in a large number of contexts, including Internet advertising (Rayo and Segal 2010), organizational communication (Jehiel 2014), financial regulation (Goldstein and Leitner 2015), medical testing (Schweizer and Szech 2018), medical research design (Kolotilin 2013; Au 2015), government control of the media (Gehlbach and Sonin 2014;

Kolotilin et.al. 2017), entertainment (Ely et al. 2015), and reporting of academic performance (Au and Kawai 2017a; Wu 2018).

Experimental studies on Bayesian persuasion are relatively scant. Frechette, Lizzeri, and Perego (2017) propose a unified experimental framework to test communications via cheap-talk, disclosure of verifiable information, and Bayesian persuasion. Concerning the tests for Bayesian persuasion model, the design they adopt is very different from ours. First, while we ask subjects who play the role of the sender to choose a distribution over posteriors, they ask subjects to choose a conditional distribution of signals. Second, they restrict the message space to three, whereas we give the senders more flexibility and allow for up to 5 messages. An advantage of our approach is that the receiver in our game faces a very simple decision-problem that does not involve any probability updating: they are given the exact posterior. Therefore, mistakes in probability calculation or non-Bayesian updating are immediately ruled out as explanations to any behavioral responses by the receiver. Moreover, the flexibility in the design of information structure (ball allocation) allows us to test whether the prediction of using only two signals holds in the laboratory. Nguyen (2017) conducts laboratory experiments of Bayesian persuasion games but with a much restricted space of feasible information structures. She finds that subjects are able to choose the optimal information structure given sufficient experience and feedback.

Our Bayesian persuasion setting is somewhat similar to an ultimatum game, as the sender has "full bargaining power" in the sense that she can choose an information structure that leaves little rent to the receiver. Ultimatum games have been first studied experimentally in Güth, Schmittberger, and Schwarze (1982), which has triggered a large number of follow-up studies (see Güth and Kocher (2014) for a recent survey). A robust finding in these studies is that when the proposer only offers a very small amount (say 1% of the pie) for the responder, most responders reciprocate negatively by rejecting the offer, even though classical economic theory predicts that the responder should accept (e.g., Kagel, Kim, and Moserthat,1996; and Roth, 2005). In a related vein, intentional reciprocation has also been documented in studies of gift-exchange games, such as Charness and Haruvy (2002), and Charness and Levine (2007). We find that similar forces can be at work in a persuasion setting. Choosing an information structure that is too opaque leaves too little rent to the receiver, who may punish the sender for being too "stingy" by refusing to be persuaded. On the other hand, offering a transparent disclosure benefits the receiver, who is then more likely to reciprocate by acting favorably to the sender.

Models of reciprocal behaviors have been proposed by Falk and Fischbacher (2006), Rabin (1993), and Fehr and Schmidt (1999). Common to these approaches is that people are assumed to be willing to

sacrifice their own monetary payoffs for a more desirable payoff allocation (among all players).[1] In Falk and Fischbacher (2006), players evaluate the kindness of an action by interpersonal payoff comparison and reward (punish) a kind (unkind) action. In Rabin (1993), players evaluate kindness by comparing the received payoff with an equitable payoff (determined by the average of the highest and the lowest possible payoffs).[2] In Fehr and Schmidt (1999), players are willing to take costly actions to reduce payoff inequality. In this paper, we follow Falk and Fischbacher (2006) in modelling reciprocity, and we will discuss implications of using alternative modelling approaches in our setting at the end of Section 2.

## 2. A Model of Bayesian Persuasion with Reciprocity

Our theory of reciprocation in persuasion is developed in this section. In Section 2.1, we begin with a quick review of the Bayesian persuasion model of Kamenica and Gentzkow (2011). We then specify the particular persuasion setting that we study. In Section 2.2, reciprocity is introduced into the model, and the reciprocity equilibrium is defined and computed. We conclude this section by developing and explicitly stating our hypotheses.

### 2.1 Bayesian Persuasion

We briefly outline the Bayesian persuasion model of Kamenica and Gentzkow (2011). Let's begin with the simplest setting of one (male) sender and one (female) receiver. In the model, the objective of the sender is to persuade the receiver to take a certain course of action. To this end, the sender can design the signal structure through which the receiver can learn about an underlying state of the world affecting her payoff of taking the action. The model imposes no constraint on the set of signal structures the sender can choose from. A conflict of interest arises if the payoff function of the sender does not coincide with that of the receiver. Kamenica and Gentzkow (2011) show that the signal structure design problem is equivalent to choosing a distribution of posterior beliefs that respects Bayes rule.

Formally, denote the state of the world by $\omega \in \Omega$ and the prior distribution over states by $\mu \in \Delta(\Omega)$. The payoffs of the sender and the receiver are respectively denoted by $v(\alpha, \omega)$ and $u(\alpha, \omega)$, where $\alpha \in A$ is the action chosen by the receiver. A signal structure on $\omega$ consists of a signal space M and a conditional distribution function $f : M \times \Omega \to [0, 1]$. Kamenica and Gentzkow (2011) show that it is without loss of generality to assume the set of available strategies to the sender is the set of Bayes-plausible distributions of posterior beliefs, i.e., $\{\mu \in \Delta(\Delta(\Omega)) : E_\mu[\omega] = \mu\}$.

[1] See Charness and Rabin (2002) for comparing and testing these theories in a unified framework.
[2] Dufwenberg and Kirchsteiger (1998) extend Rabin's model to sequential games.

We investigate to the following persuasion game. There are two possible states of the world, red and black, with some prior distribution. The receiver has two available actions, R and B, and she gets a positive payoff (normalized to one throughout this section) if and only if her action matches with the state (i.e., choosing R when the state is red, and B when the state is black). The sender, on the other hand, gets a positive payoff (again normalized to one) if and only if the receiver chooses R. The objective of the sender is thus to maximize the chance that the receiver, after learning some extra information on the state of the world, is willing to choose R. In the notations above, $\Omega = \{\text{red}, \text{black}\}$, and $A = \{R, B\}$. As the state is binary, we can represent any belief over the state by the probability that the state is red. With this convention, the prior belief $\mu$ is a scalar specifying the prior probability that the state is red. Moreover, the sender's payoff has $v(\alpha, \omega) = 1$ if and only if $\alpha = R$, whereas the receiver's payoff has $u(\alpha, \omega) = 1$ if and only if either $\alpha = R$ and $\omega = red$, or $\alpha = B$ and $\omega = black$.

If he believes that the receiver is an expected-utility maximizer (as in the standard Bayesian persuasion model), then the optimal signal structure consists of only two signal realizations, one that makes the receiver indifferent between R and B, another that fully reveals the state is black. In the notations above, $|\mathrm{M}| = 2$. If we denote the two signals by h and l respectively, then the first signal $h$ leads to a posterior belief $\Pr(\omega = \text{red}|\text{h}) = 0.5$ and the second signal $l$ gives a posterior $\Pr(\omega = \text{red}|\text{l}) = 0$.

The experiment we design fits the setting above. However, instead of framing the problem as one of choosing a distribution of posterior beliefs, we ask subjects who play the role of senders to allocate colored balls into a number of urns. One of the urns will be drawn randomly, the probability of which depends on the total number of balls in the urn. Upon learning the composition of the drawn urn, subjects who play the role of receivers then choose between red and black. Finally, one ball is drawn from the urn randomly, and the receivers collect a positive payoff if and only if the color of the ball coincides with their choice. We will explain in greater details why the game of Bayesian persuasion and the game of ball allocation are strategically equivalent in Section 3. We conclude this subsection with the following lemma that states the predictions of the standard Bayesian persuasion model for the game we consider.

**Lemma 1** The unique perfect Bayesian equilibrium is as follows. The receiver chooses R if and only if the posterior belief induced the signal realization is no less than 0.5. If $\mu < \frac{1}{2}$, the sender chooses a signal structure that induces only two posteriors 0 and 0.5. If $\mu = \frac{1}{2}$, the sender chooses a signal structure that induces only a posterior 0.5.

For later reference, note that in the unique perfect Bayesian equilibrium, the receiver's response, given by a simple step function in posterior beliefs, is independent of the prior $\mu$. Moreover, provided that $\mu \leq \frac{1}{2}$, only posteriors 0 and 0.5 would be induced by the sender. As shown below, the predictions could be very different once reciprocal preference is introduced.

## 2.2 Persuasion with Reciprocity

In this subsection, we introduce reciprocity into the Bayesian persuasion game we studied following the approach of Falk and Fischbacher (2006). Specifically, on top of the standard monetary payoffs, a player's utility function consists of terms that capture the kindness of the other player, as well as an appropriate reciprocation to the received treatment. Let $\pi_i(\sigma_S, \sigma_R)$ be the monetary payoff of player $i \in \{S, R\}$ if the sender and the receiver play strategy $\sigma_S$ and $\sigma_R$ respectively. Consider first the receiver's utility function. Let $k_i(\sigma'_R, \sigma''_R)$ be the sender's kindness perceived by the receiver, where $\sigma'_R$ is the receiver's first-order belief concerning the sender's strategy, and $\sigma''_R$ is the receiver's second-order belief concerning the sender's belief on the receiver's strategy. Falk and Fischbacher (2006) specify that

$$k_S(\sigma'_R, \sigma''_R) = \pi_R(\sigma'_R, \sigma''_R) - \pi_S(\sigma'_R, \sigma''_R).$$

That is, the higher the payoff that the sender's strategy is perceived to bring to the receiver relative to the sender's own payoff, the kinder the receiver thinks the sender is. Next, the reciprocation term of the receiver is defined as

$$\rho_R(\sigma_R, \sigma'_R, \sigma''_R) = \pi_S(\sigma'_R, \sigma_R) - \pi_S(\sigma'_R, \sigma''_R).$$

The reciprocation term above can be interpreted as the alteration in the sender's payoff brought about by the receiver changing his/her strategy from $\sigma''_R$ to $\sigma_R$. The receiver's reciprocity utility is defined as the product of the kindness term $k_S(\sigma'_R, \sigma''_R)$ and the reciprocation term $\rho_R(\sigma_R, \sigma'_R, \sigma''_R)$. The idea is that if the receiver perceives the sender to be kind, then he/she derives a positive utility by returning the sender a favor. Conversely, if the receiver perceives the sender to be unkind, then he/she derives a positive utility by taking action to lower the sender's monetary payoff. The receiver's overall utility is defined as

$$U_R(\sigma_S, \sigma_R; \sigma'_R, \sigma''_R) \equiv \pi_R(\sigma_S, \sigma_R) + \lambda_R k_S(\sigma'_R, \sigma''_R)\rho_R(\sigma_R, \sigma'_R, \sigma''_R),$$

where $\lambda_R \leq 1$ is the receiver's reciprocation parameter. It is a positive constant that measures the strength of the reciprocal preference.

While we can formulate the sender's reciprocal utility in a similar manner, we proceed our analysis by assuming that the sender does not care about reciprocation. In other words, we assume that the sender's reciprocation parameter is zero and she maximizes her monetary payoff only. This simplification is without loss of generality if the sender's reciprocation parameter is not too large. As shown in Falk and Fischbacher (2006), the proposer's reciprocation incentives matter in an ultimatum game if and only if the proposer's reciprocation parameter is significantly larger than that of the responder. As the sender in our Bayesian persuasion game plays a role similar to a proposer in an ultimatum game, a similar result holds in our setting.

As the players' utilities are assumed to depend on their beliefs, the reciprocity game proposed by Falk and Fischbacher (2006) belongs to the class of psychological games pioneered by Geanakoplos, Pearce, and Stacchetti (1989). The equilibrium notion we adopt is standard in psychological games. First, given beliefs, each player maximizes their expected utility. Second, the beliefs match the actual behaviors.

We specify the kindness terms in our Bayesian persuasion game below. Denote by μ the initial fraction of red balls. For simplicity, we make the following restrictions on players' strategies. Each sender provides at most two urns. Feasibility implies that one of which has a fraction of red balls no higher than μ, whereas the other has a fraction of red balls no lower than μ.[3] Under this restriction of strategy space, a generic strategy of the sender is a pair of fractions $(p, q)$, with $p \geq \mu \geq q$, describing the respective proportion of red balls in the two urns designed. A generic strategy of the receiver specifies the probability $\sigma_R(p)$ of guessing red after an urn with a fraction p of red balls is drawn.

We assume that given a certain equilibrium sender strategy $(p^*, q^*)$, the receiver evaluates the sender's kindness only by the fraction of red balls in the drawn urn. Specifically, suppose the fraction of red balls is $p \geq \mu$ in the drawn urn, and let $\sigma''_R$ be the receiver's second-order belief about his own strategy. Then the sender's kindness perceived by the receiver is given by

$$
\begin{aligned}
&k_S\left(p, \sigma''_R; (p^*, q^*)\right) \\
&\qquad = \left[\begin{array}{l} \dfrac{p-\mu}{p-q^*}\left(q^*\sigma''_R(q^*) + (1-q^*)\left(1-\sigma''_R(q^*)\right)\right) \\ \quad + \dfrac{\mu-q^*}{p-q^*}\left(p\sigma''_R(p) + (1-p)(1-\sigma''_R(p))\right) \end{array}\right] \\
&\qquad - \left[\frac{p-\mu}{p-q^*}\sigma''_R(q^*) + \frac{\mu-q^*}{p-q^*}\sigma''_R(p)\right].
\end{aligned}
$$

[3] In our experiments, it occurred more than 65% of the time that no more than two effective urns (in term of the fraction of red ball in the urn) were chosen by senders.

Note that in the specification above, even though the receiver does not observe the composition of the undrawn urn, he assumes that its fraction of red balls stays at the equilibrium value of $q^*$. To understand the expression above, note that if the sender's strategy is $(p, q^*)$, then with probability $\frac{p-\mu}{p-q^*}$, the urn with fraction $q^* \leq \mu$ of red balls is drawn. In this case, the payoffs of the receiver and the sender are $q^*\sigma''_R(q^*) + (1-q^*)\big(1-\sigma''_R(q^*)\big)$ and $\sigma''_R(q^*)$ respectively. With complementary probability $\frac{\mu-q^*}{p-q^*}$, the urn with a fraction $p \geq \mu$ of red balls is drawn. In this case, the expected payoffs of the receiver and the sender are $(p\sigma''_R(p) + (1-p)(1-\sigma''_R(p)))$ and $\sigma''_R(p)$ respectively.

Similarly, if the drawn urn has a fraction $q \leq \mu$ of red balls, the sender's perceived kindness is

$$k_S\big(q, \sigma''_R; (p^*, q^*)\big) = \left[\begin{array}{l} \frac{p^*-\mu}{p^*-q}\Big(q\sigma''_R(q) + (1-q)\big(1-\sigma''_R(q)\big)\Big) \\ +\frac{\mu-q}{p-q}(p^*\sigma''_R(p^*) + (1-p^*)(1-\sigma''_R(p^*)))\end{array}\right] - \left[\frac{p^*-\mu}{p^*-q}\sigma''_R(q) + \frac{\mu-q}{p^*-q}\sigma''_R(p^*)\right].$$

In sum, a pair of strategy $((p^*, q^*), \sigma_R(\cdot))$ constitutes a reciprocity equilibrium if and only if

(i) the sender's strategy maximizes her utility given belief $\sigma_R(\cdot)$, i.e.,

$$(p^*, q^*) \in argmax_{\{(p',q'):p' \geq \mu \geq q'\}} \frac{p'-\mu}{p'-q'}\sigma_R(q') + \frac{\mu-q'}{p'-q'}\sigma_R(p'); \text{ and}$$

(ii.a) if an urn with a fraction $p \geq \mu$ of red balls realized, the receiver maximizes her utility given beliefs $(q^*, \sigma_R(\cdot))$, i.e.,

$$\sigma_R(p) \in argmax_{\sigma \in [0,1]} p\sigma + (1-p)(1-\sigma) + \lambda_R k_S\big(p, \sigma_R; (p^*, q^*)\big)[\sigma - \sigma_R(p)]; \text{ and}$$

(ii.b) if an urn with a fraction $q \leq \mu$ of red balls realized, the receiver maximizes her utility given beliefs $(p^*, \sigma_R(\cdot))$,, i.e.,

$$\sigma_R(q) \in argmax_{\sigma \in [0,1]} q\sigma + (1-q)(1-\sigma) + \lambda_R k_S\big(q, \sigma_R; (p^*, q^*)\big)[\sigma - \sigma_R(q)].$$

The proposition below states the reciprocity equilibrium.

**Proposition 1** Let $\mu \leq \frac{1}{2}$. The reciprocity equilibrium (in which the sender offers two urns) is unique. The sender chooses $(p^*, 0)$ where $p^* = \frac{1}{4}[(1-\lambda_R(1+\mu)) + \sqrt{((1-\lambda_R(1+\mu))^2 + 16\mu\lambda_R}]$. Upon observing an urn with a fraction p of red balls, the receiver's probability of guessing red is given by

$$\sigma_R(p) = \begin{cases} \max\left\{0, \dfrac{\frac{(2\mathrm{p}-1)(p^*-\mathrm{p})}{\lambda_R} + (1-\mu)(p^*-\mathrm{p}) - 2(\mu-\mathrm{p})(1-p^*)}{2(p^*-\mu)(1-\mathrm{p})}\right\} & if\ p < \mu \\ \min\left\{\dfrac{\frac{(2\mathrm{p}-1)\mathrm{p}}{\lambda_R} + (1-\mu)\mathrm{p}}{2\mu(1-\mathrm{p})}, 1\right\} & if\ p \geq \mu \end{cases}.$$

The proof of the proposition can be found in the appendix. The proposition implies that the receiver's probability of guessing red $\sigma_R(p)$ is not a step-function in the fraction p of red balls of the drawn urn, as predicted in the standard Bayesian persuasion model in which the receiver is assumed to maximize expected payoff (recall Lemma 1). If the receiver's reciprocity parameter $\lambda_R$ is positive, $\sigma_R(p)$ increases continuously over an interval of $p$. Moreover, if $\lambda_R$ is sufficiently large, $\sigma_R(p)$ can be positive even if the fraction $p$ of red balls is less than ½, and it can be less than one even if the fraction $p$ of red balls exceeds ½. Figure 1 below illustrates the receiver's equilibrium strategy $\sigma_R(p)$ for the case $\lambda_R = 0.6$.

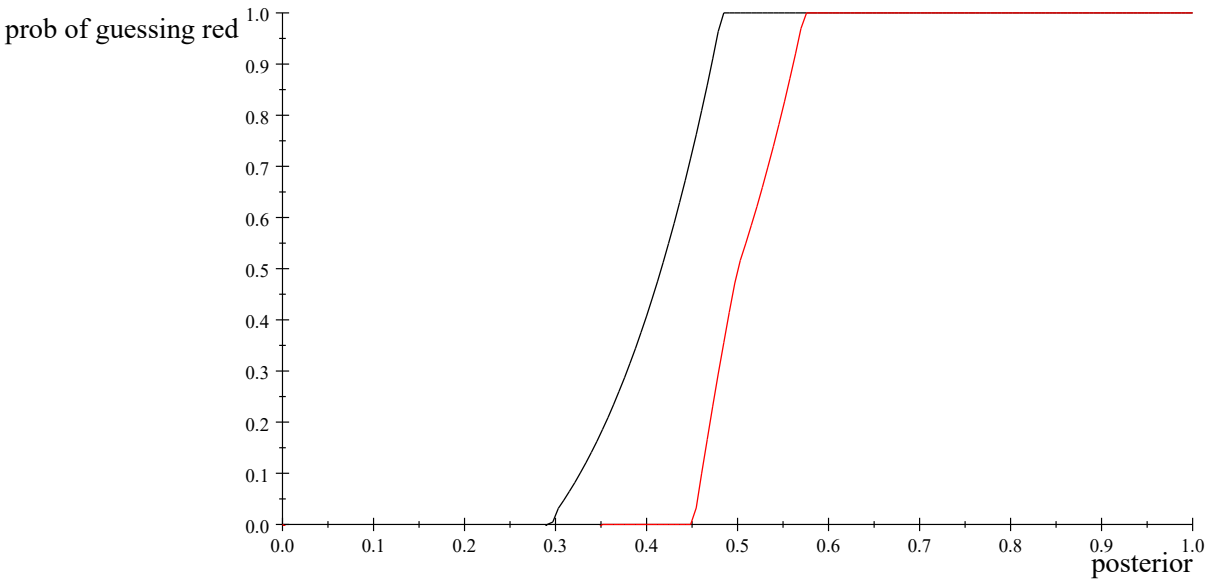


Figure 1: Receiver's equilibrium responses (Black curve: $\mu = 0.3$; red curve: $\mu = 0.5$)

The discussion above leads to our first hypothesis.

**Hypothesis 1** The probability that the receiver guesses red increases continuously in the fraction of red balls in the drawn urn around the 50% mark.

Proposition 1 implies that an increase in $\mu$ would shift $\sigma_R(p)$ downwards. Moreover, as $\mu$ increases, the sender's equilibrium strategy $(p^*, 0)$ would have $p^*$ going up. In other words, the sender provides a more "generous" ball allocation, with the urn that contains a positive number of red balls having a higher proportion of red balls. This corresponds to a more transparent information disclosure.

**Corollary 1** Suppose $\mu < \mu' \leq \frac{1}{2}$. The receiver's probability of guessing red $\sigma_R(p)$ is weakly lower when the initial fraction of red balls is $\mu'$ than when the fraction is $\mu$. Moreover, in the sender's strategy $(p^*, 0)$, $p^*$ goes up with initial fraction of red balls, i.e., the urn that contains a positive number of red balls has a higher proportion of red balls when the initial fraction is $\mu'$ than when it is $\mu$.

In Figure 1, it is apparent that $\sigma_R(p)$ is weakly lower for the case $\mu = 0.5$ than the case $\mu = 0.3$, and strictly so when $p$ is between 0.3 and 0.55. Moreover, when $\mu = 0.3$, the sender's optimal strategy has a $p^*$ close to 0.5; and when $\mu = 0.5$, the sender's optimal strategy has a $p^*$ close to 0.6. Figure 2 plots the sender's equilibrium choice of $p^*$ against $\lambda_R$. It is clear that the sender's choice of $p^*$ is higher for every level of $\lambda_R$.

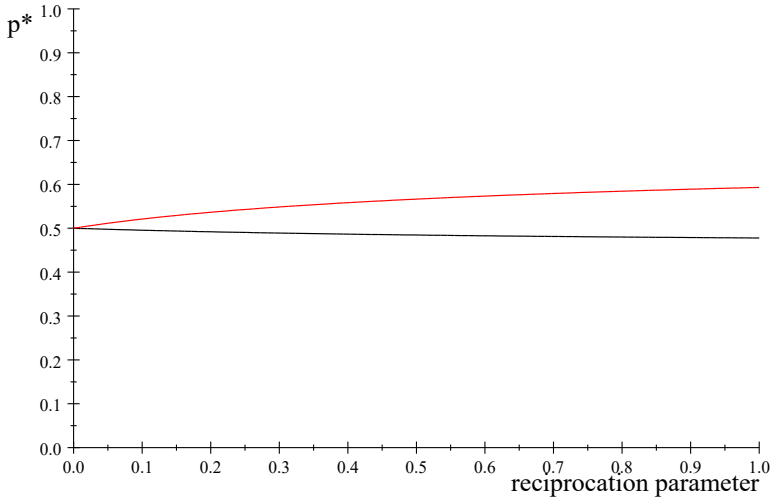


Figure 2: Sender's Equilibrium Choice of $p^*$ against $\lambda_R$ (Black curve: $\mu = 0.3$; red curve: $\mu = 0.5$)

The intuition of Corollary 1 is analogous to that of an ultimatum game. The receiver demands a "fair share" of the surplus and is therefore willing to sacrifice his own monetary payoffs to punish the sender

if she perceives the sender to be unkind. In an ultimatum game, the punishment takes the form of rejecting the proposer's offer (thus giving zero payoff to both players), whereas in the Bayesian persuasion game here, the punishment takes the form of guessing black (thus lowering his own expected payoff in order to make the sender getting a zero payoff). When the initial fraction of red balls is higher, the sender has a higher expected payoff holding fixed the receiver's behavior. As a result, for each urn drawn, the receiver would perceive the sender to be less kind than when the initial fraction of red balls is lower, and consequently, they lower the probability of guessing red. This leads to our second hypothesis.

**Hypothesis 2** An increase in the initial fraction of red balls weakly lowers the probability that the receiver guesses red for any urn composition.

According to the second part of Corollary 1, as an increase in the initial number of red balls makes the receiver more demanding, the sender finds it optimal to design urns with more extreme composition, as a higher fraction of red balls is needed to induce a high probability that the receiver guesses red. This is our third hypothesis.

**Hypothesis 3** When the initial number of red balls increases, the sender makes the fraction of red balls in the urns more extreme (i.e., closer to 0 and 1).

Recall Lemma 1 states that in the absence of any reciprocal preference, the prior belief, which corresponds to the initial fraction of red balls here, affects neither the receiver's response function, nor the "good signal" (which corresponds to the fraction of red balls in the urn that contains a positive number of red balls). In other words, the standard Bayesian persuasion model without reciprocation incentives, would predict, in sharp contrast to Hypotheses 2 and 3 above, that neither the receiver's strategy $\sigma_R(p)$ nor the sender's strategy $(p^*, 0)$ would be affected by the prior belief.

Our last hypothesis concerns the expected payoff of the players. While it is intuitive that the sender would benefit from an increase in the initial fraction of red balls (a more favourable common prior belief), a priori, it is less clear whether the receiver would benefit. In fact, in a Bayesian persuasion model without reciprocity, an increase in the prior could strictly harm the receiver. To see this, note that the receiver gets the positive payoff for certain if the drawn urn has 0% red ball, and gets the positive payoff with probability 50% if the drawn urn has 50% red ball. When the initial fraction of red balls is $\mu < \frac{1}{2}$, the receiver's ex-ante probability of getting a positive payoff is $\frac{\mu}{0.5} \times 0.5 + \left(1 - \frac{\mu}{0.5}\right) \times 1 = 1 - \mu$, which is strictly decreasing in $\mu$. The following corollary states that, however, when the reciprocation parameter is sufficiently large, the receiver's expected payoff can become increasing in $\mu$.

**Corollary 2** Suppose $\mu < \mu' \leq \frac{1}{2}$.. The sender's expected monetary payoff is higher if the initial fraction of red balls is $\mu'$ than when it is $\mu$. The same is true for the receiver's expected monetary payoff if the reciprocation parameter $\lambda_R$ is sufficiently large.

The proof can be found in the appendix. The corollary leads to the following hypothesis on the expected monetary payoffs of the players.

**Hypothesis 4** When the initial number of red balls increases, both the expected monetary payoffs of both sender and the receiver increase.

We conclude this section by briefly discussing alternative common approaches of modelling reciprocity in the literature. A notable feature of our model is that the receiver evaluates the sender's kindness by comparing the sender's expected payoff with hers. In Rabin (1993), kindness is evaluated by comparing the actual payoff with an equitable payoff (that is determined by the average of the highest and lowest possible payoffs). In our setting, an increase in the initial fraction of red balls from $\mu$ to some $\mu' \leq \frac{1}{2}$, holding other things constant, would actually decrease the receiver's lowest possible payoff (without changing the highest possible payoff), so the equitable payoff to the receiver goes down. Any strategy of the sender would then be perceived as more kind, and the receiver would be more willing to guess red. Consequently, the predictions in Hypotheses 2 to 4 would be reversed under Rabin (1993)'s formulation. Fehr and Schmidt (1999) take a consequentialist approach and assume that players take actions to reduce payoff inequality between the players. As the sender's intention does not matter in determining payoff inequality, only ex-post payoff-relevant information enters into players' utility function. Therefore, in our setting, the approach of Fehr and Schmidt (1999) would predict that the initial fraction of red balls has no effect on the receiver's responses and the sender's strategy.[4]

## 3. Experimental Design

In this section, we first describe our experiment procedures in details in Section 3.1. We then explain why our design is essentially strategically equivalent to a Bayesian persuasion setting in Section 3.2.

### 3.1 Experiment Procedures

In our experiment, there were 10 rounds of decision making. In each round, subjects were randomly matched into pairs. In each matched pair, one subject was assigned to the role of Player A; and the other

[4] More precisely, the percentages of red balls in the constructed urns would remain unchanged, though the numbers of balls in the constructed urns would differ.

the role of Player B. Each subject's role remained fixed throughout the experiment. Subjects' decisions were anonymous. The game had two stages.

### Stage 1

In each round, player A were asked to allocate (all of) 100 colored balls into 5 urns. There are two treatments. In the 30-red-70-black treatment, there were 30 red balls and 70 black balls. In 50-red-50-black treatment, there were 50 red balls and 50 black balls. Player A was asked to fill in the following table:

| Urn | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Red | | | | | |
| Black | | | | | |

Player A had to determine the number of red balls and black balls in each of the urns. An empty urn is allowed. All 100 balls must be allocated to one of the urns; in other words, the sum of balls in the urns must be 100.

Then, one urn would be randomly drawn. The total number of balls in each urn would determine the probability that the urn will be drawn. For example, if urn 1 contained 80 balls, then this urn would be drawn will probability 0.8. Player A could check the total number of balls allocated by clicking the icon "Check" in the program before submitting his/her decision.

### Stage 2

After an urn was drawn, Player B received a message on the screen specifying the number of red balls and number of black balls in the urn. For example:

*The urn contains 20 red balls and 20 black balls.*

Then, a ball would be randomly drawn from the urn, and player B was asked to guess whether the ball drawn is red or black.

### Payoff

Player A received 40 Hong Kong dollars if player B guessed that the ball was red, regardless of the actual color of the ball drawn. Player A received 0 dollars if player B guessed that the ball was black,

regardless of the actual color of the ball drawn. In the end of the experiment, one round would be randomly drawn for payment.

If player B's guess coincided with the ball drawn, he/she received 40 dollars. If the guess was incorrect, his/her payoff was zero.

### Information Feedback

At the end of each round, subjects were informed about (i) the urn drawn, (ii) the guess made by Player B, (iii) the color of the ball drawn, and (iv) earning.

In total, 162 subjects participated in 8 sessions of the experiment (4 sessions for each treatment). Each subject participated in only one session. The subjects were undergraduate students in a major university in Hong Kong, and they were randomly recruited using an electronic recruitment system. The experiment was conducted using the program z-tree (Fischbacher, 2007) and took place in a laboratory where subjects were randomly seated in partitioned cubicles. Before the beginning of the 10 decision-making rounds, subjects were given one practice round. Out of the 10 rounds, one round was randomly drawn for payment. In the end of the experiments, subjects completed a questionnaire in which they were asked the reasons behind their choices in the experiment. Subjects received a show-up fee of 40 Hong Kong dollars, in addition to earnings from the randomly-drawn round.

## 3.2 Strategic Equivalence between Our Experiment and Bayesian Persuasion

While subjects who play the role of Player A in our experiment chooses a ball allocation among urns, their decision problem is strategically equivalent to choosing a distribution of posterior. There is a one-to-one mapping between the ball allocation problem and that of signal structure design. First, the colors of the balls correspond to the states, and the initial proportion of red balls corresponds to the prior probability distribution (0.3 in the 30-red-70-black treatment and 0.5 in the 50-red-50-black treatment). Second, each urn corresponds to one possible signal realization. Here, an urn with more assigned balls corresponds to a signal with a higher likelihood of materialization. Moreover, an urn with a larger proportion of red balls corresponds to a signal that is more indicative the state being red. It is straightforward to show that if there are infinitely many available urns, and if the balls are infinitely divisible, the ball-allocation problem is mathematically equivalent to the signal-structure design problem.

To see this formally, given a prior distribution $\mu_0$ and a signal structure $f(s|\omega)$, the posterior probability that the state is red conditional on signal realization $s_0$ is $\frac{\mu_0(red)f(s_0|red)}{\mu_0(red)f(s_0|red)+\mu_0(black)f(s_0|black)}$, and the probability that signal realization $s_0$ materializes is $\mu_0(red)f(s_0|red) + \mu_0(black)f(s_0|black)$. If we

interpret the prior distribution $\mu_0$ as a collection of probability masses, with $\mu_0(red)$ units of *red* mass and $\mu_0(black)$ units of *black* mass, then the signal realization $s_0$ can be interpreted as an "urn" that contains a subset of these probability masses; specifically it contains a probability mass of $\mu_0(red)f(s_0|red) + \mu_0(black)f(s_0|black)$ units, out of which $\mu_0(red)f(s_0|red) + \mu_0$ units are *red* mass. Upon observing signal $s_0$, the receiver's posterior belief about state being red is $\frac{\mu_0(red)f(s_0|red)}{\mu_0(red)f(s_0|red)+\mu_0(black)f(s_0|black)}$; this probability is identical to that of drawing a unit of red mass out of $\mu_0(red)f(s_0|red) + \mu_0(black)f(s_0|black)$ units of probability masses.

For practical implementation, we restrict the maximum number of urns to 5 and provide subjects with 100 indivisible balls. It is worth noting that even with these restrictions, the optimal ball allocation (or equivalently, the optimal signal structure) in the absence of the finiteness constraint can still be implemented in our setting. Specifically, in the 30-red-70-black treatment, an optimal allocation involves putting all balls into two urns: one urn consists of 30 red balls and 30 black balls; whereas the other urn consists of 40 black balls. In the 50-red-50-black treatment, an optimal allocation involves putting all balls in a single urn.

## 4. Experimental Results

We report our experimental results in this section. Section 4.1 discusses findings about the responses of the receiver (Player B), and Section 4.2 discusses the ball-allocation choices of the senders (Player A).

### 4.1 Receivers' Responses

Figure 3 reports Player Bs' average frequency of guessing red when presented with urns of different compositions for both the 30-red-70black treatment and the 50-red-50-black treatment. It is clear that in both treatments, Player Bs almost never guesses red when the urn has less than 30% of red balls, and they always guess red when the urn has more than 70% of red balls. Interestingly, the probability that Player B guesses red is significantly positive when presented with urns with 40-50% (i.e., at least 40% and lower than 50%) of red balls, even though such a choice gives her a negative expected payoff. More specifically, in the 30-red-70-black treatment, the probability of guessing red is 0.39 which is significantly higher than zero, with p-value equal to 0.00 under the one-sample t-test. In the 50-red-50-black treatment, the probability of guessing red is 0.19, which is also significantly higher than zero, with p-value equal to 0.00. The probability of guessing red spikes once the 50% mark is hit, though it falls short of 100% by a large margin. In the 30-red-70-black treatment, when presented with urns with 50-52% of red balls, Player B guesses red with a probability of 0.56 only. When presented with urns with 52-54% of red balls, Player B guesses red with a probability of 0.7, which is significantly lower

than 1, with p-value equals to 0.00 under the one-sample t-test. These probabilities are even lower in the 50-red-50-black treatment.

These observations indicate that Player Bs' behaviors are not completely consistent with expected-utility maximization, which would call for guessing red whenever the proportion of red balls exceeds 50%. It also stands in sharp contrast to the prediction of the standard Bayesian persuasion model that Player B's response is a step function (equals 0 if the fraction is less than 50% and equals 1 if the fraction exceeds 50%).

**Result 1:** Player B does not guess red for sure even if the urn contains more than 50% of red ball. For urns with fractions of red balls between 50% and 70%, the probability that player B guesses red is significantly lower than 1.

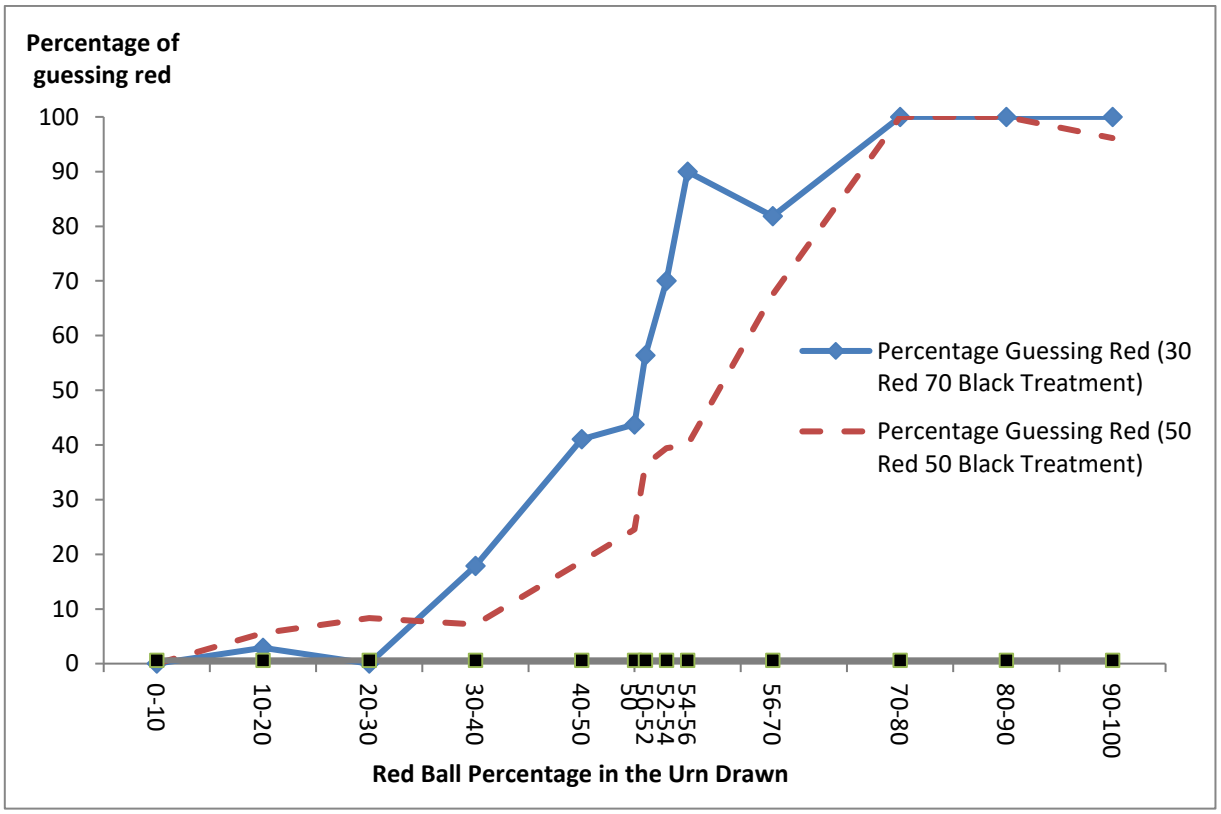


Figure 3: Percentage of guessing red against urn composition

Instead of being a step-function as predicted by standard Bayesian persuasion model, Figure 3 shows that Player Bs' responses change continuously in the fraction of red balls in the urn. In fact, as the fraction of red balls increases from 30% to 70%, Player Bs' probability of guessing red strictly

increases in a continuous manner. Column 1 of Table 1 reports the marginal effect coefficient estimations of the probability of guessing red. The independent variables are the percentage of red balls in the drawn urn, and a dummy variable for the 50-red-50-black treatment. It is confirmed that the higher is the percentage of red balls, the more likely that player B will guess red. Column 2 of Table 1 reports the same regression for the subsample where the percentage of red ball in the drawn urn is higher than 50%. It is also found that the probability of guessing red increases with the percentage of red balls. Note that Bayesian persuasion predicts that the coefficient to be zero. Hence, Hypothesis 1 is supported.

**Result 2:** As the percentage of the red balls in the drawn urn increases from 30% to 70%, the probability of guessing red by Player B increases continuously.

Table 1. Determinants of Probability of Guessing Red

| | Dependent variable: Guessing Red | |
|---|---|---|
| | (1) Full Sample | (2) More than 50 Percentage of Red Ball in the Drawn Urn |
| Percentage of Red Ball | 0.02*** | 0.02*** |
| | (0.001) | (0.01) |
| 50-red-50-black Treatment | -0.14*** | -0.16*** |
| | (0.03) | (0.17) |
| Number of Observations | 810 | 349 |
| Pseudo R-square | 0.50 | 0.18 |

*Notes:* This table reports the marginal effect coefficient estimates of Probit regression. The independent variable Percentage of Red Ball is the percentage of red ball in the drawn urn. The independent variable 50-black-50-red treatment is the dummy which equals to 1 for the 50-red-50-black treatment, zero otherwise. *, **, *** denotes significance at the 10%, 5%, and 1% levels, respectively.

In the Bayesian persuasion treatment[5] of the experiments conducted in Frechette, Lizzeri, and Perego (2017), they also find that the receivers' (Player B) responses are not a step function. Rather, it is a linear function over the whole interval from 0% to 100%. In contrast, Figure 3 shows that in our experiments, Player Bs' responses are not linear in the fraction of red balls. Instead, the probability of

[5] See the subfigure for Treatment V100 in Figure 10 in Frechette, Lizzeri, and Perego (2017).

guessing red is flat in the interval from 0% to 30% and in the interval 70% to 100%, and strictly increases only in the 30%-70% range. This pattern is closer to the prediction of the model with reciprocity analyzed in Section 2 (see Figure 1).

We now proceed to compare player B's responses in the two treatments. Figure 3 shows that given the same fraction of red balls in the drawn urn in the range of 30-70%, the probability that player B guesses red is strictly lower in the 50-red-50-black treatment than the 30-red-70-black treatment for almost all fractions of red balls in the drawn urn. Moreover, as shown in Table 1, after controlling for the percentage of the red balls, player Bs in the 50-red-50-black treatments are significantly less likely to guess red. We also run additional regressions for the respective subsamples of urns with red balls exceeding 70 percent, and of urns with red balls less than 30 percent, and find no significant treatment differences in these subsamples.

**Result 3:** An increase in the initial fraction of red balls weakly lowers the probability that the receiver guesses red for any urn composition, and strictly so when the fraction of red balls is in the range of 30-70%.

This result supports Hypothesis 2. In fact, this finding is very intuitive from the perspective of reciprocation. Consider, for example, the case of an urn with 50% of red ball is drawn. Facing such an urn, player B in the 30-red-70-black treatment may interpret that player B is relatively kind (as the percentage of red balls exceeds 30%) than the case of 50-red-50-black treatment where the percentage of red ball is the same as the initial condition. Hence, player B in the 30-red-70-black treatment is more likely to guess red. Our finding echoes that of Falk, Fehr, and Fischbacher (2003). They found that in an ultimatum game, subjects exhibited different reciprocity attitudes when faced with offers that are identical in monetary term but come from different sources. In other words, reciprocity depends on not only on the absolute value of the offer, but also the inferred intent.

The subjects' responses in the questionnaire answered at the end of the experiment sessions shed interesting light on the role of reciprocation concerns in their decisions. Subjects who were assigned the role of player Bs mentioned that they would choose black when there were 50% of red balls or the percentage of red balls did not exceed the black balls by large margin. The main reason mentioned was that they did not want to receive a zero payoff while letting player B receive 40 dollars. For example, one subject wrote "When the number of the balls is similar or red balls are only 2-3 balls more, I choose black as my answer, as the probability if winning is approximately 1/2 only. Therefore, even if I am wrong, Player A cannot get his pay." Another subject wrote "If the ratio of red and black ball is 1:1. I would choose black ball so that player A can't get additional payoff." It is also clear from the responses

that player B understood that in this case, the chance to win was 50%. Hence, player B's behavior was not driven by mistaken probabilistic beliefs.

Given player B's responses differ from the prediction of standard Bayesian persuasion model, we can see that the empirically optimal ball allocation differs from the theoretically optimal urn based on the assumption that Player Bs are expected-utility maximizers (i.e., providing two urns, one of which has 50% red balls and the other urn has 0% red balls). Specifically, applying the technique of Kamenica and Gentzkow (2011) by looking for the concave closure of the payoff function in posteriors, Figure 3 reveals that in the 30-red-70-black treatment, the empirically optimal ball allocation consists of two urns, one of which has about 55% red balls, and the other has 0% red balls. The corresponding expected payoff to Player A is approximately (0.3/0.55)*0.9*40 = 19.64 dollars, significantly outperforming that of the equilibrium allocation predicted by the standard model (0.3/0.5)*0.6*40 = 14.4 dollars. The contrast is more striking in the 50-red-50-black treatment. In this case, Figure 3 reveals that the empirically optimal ball allocation consists of two urns, one of which has about 75% red balls, and the other has 0% red balls. The corresponding expected payoff to Player A is approximately (0.5/0.75)*1*40 = 26.67 dollars, significantly outperforming that of the equilibrium allocation predicted by the standard model (0.5/0.5)*0.25*40 = 10 dollars.

**Result 4:** The empirically optimal ball allocation is different from the prediction of the standard Bayesian persuasion model. In particular, the empirically optimal ball allocation involves providing an urn with a fraction of red balls significantly exceeding 50%.

## 4.2 Senders' Behaviors

Figure 4 reports the empirical frequency of urn composition chosen by Player A by pooling the ball allocations of all subjects. In either treatment, urns with 0% red balls and 50-55% red balls are the most frequently offered urns. Interestingly, urns with more than 50% red balls are offered quite often. The proportions of urns with more than 55% of red balls are about 16.83% and 39.99% in the 30-red-70-black treatment and 50-red-50-black treatment respectively. Both proportions are significantly different from zero, contradicting the prediction of the standard Bayesian persuasion model. More importantly, the proportion is significantly higher in the 50-red-50-black treatment than in the 30-red-70-black treatment, with a p-value of 0.00 under the two-sample t-test.

**Result 5**: Player As choose urns with more than 50% of red balls quite often. The empirical frequency of urns with more than 50% of red balls is much higher in the 50-red-50-black treatment than the 30-red-70-black treatment.

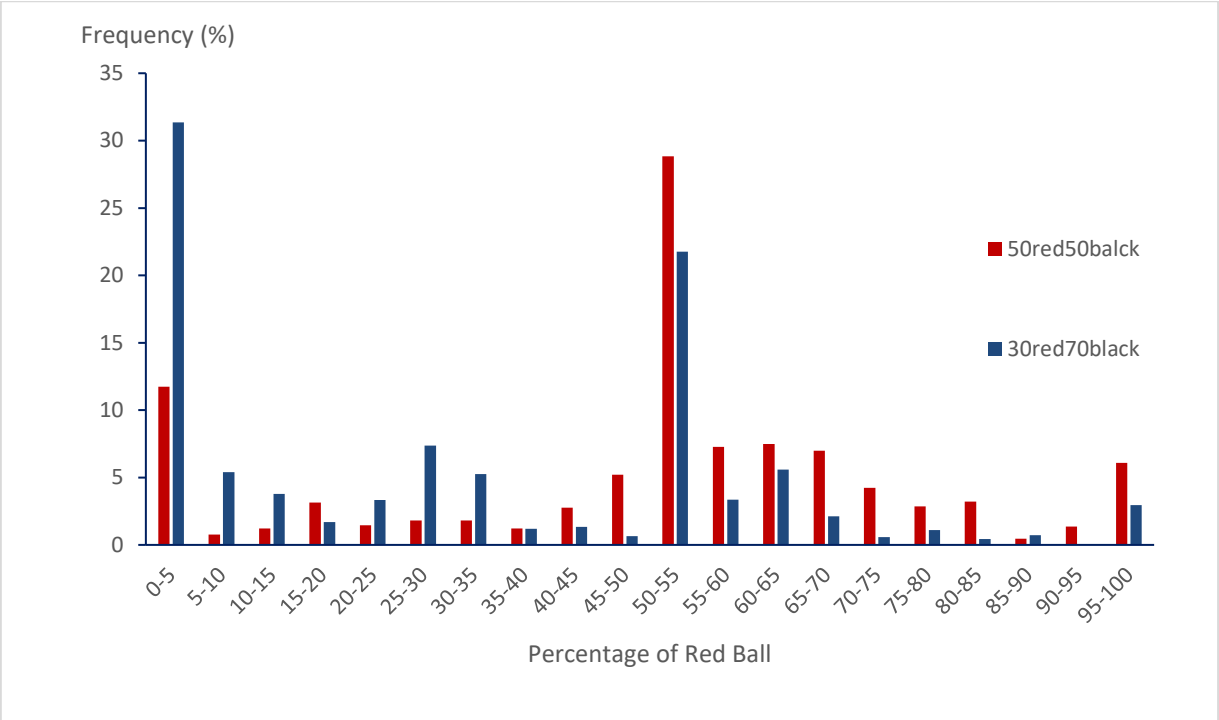


Figure 4: Distribution of Urns

Result 5 is consistent with Hypothesis 3. In particular, Player A provides urns with composition that are more favorable to Player B in the 50-red-50-black treatment. This finding indicates that Player As in our experiments respond to the difference in Player B's behavior across the two treatments.

We find that, as Player As make more "generous" offers in the 50-red-50-black treatment, Player Bs indeed benefit a lot from a higher initial fraction of red balls. The average monetary payoff of Player Bs is 25.8 dollars and 17.3 dollars in the 50-red-50-black treatment and the 30-red-70-black treatment respectively. The difference is highly significant, with a p-value of 0.00 under a two-sample t-test. Similarly, Player A also benefit from an increase in the initial fraction of red balls. Their average monetary payoffs of Player As are 16.9 dollars and 12.1 dollars in the 50-red-50-black treatment and the 30-red-70-black treatment respectively. The difference is again highly significant, with a p-value of 0.00 under a two-sample t-test. We thus have the following result.

**Result 6**: The average monetary payoff of both Player A and Player B are much higher in the 50-red-50-black treatment than the 30-red-70-black treatment.

This result supports the Hypothesis 4 that both players share the benefit of having a higher initial fraction of red balls.

We conclude with a brief discussion on the number of urns used by Player A. Recall that both the standard Bayesian persuasion model and our reciprocity model outlined in Section 2 predict that Player A uses, effectively, only two urns.[6] We find that the number of effective urns in our experiment is close to two most of the time.[7] As shown in Figure 5, Player A uses two urns most frequently in both treatments. Moreover, the frequency of using all the 5 provided urns is only about 2%, indicating that we provide subjects with sufficient flexibility in their ball-allocation choice.

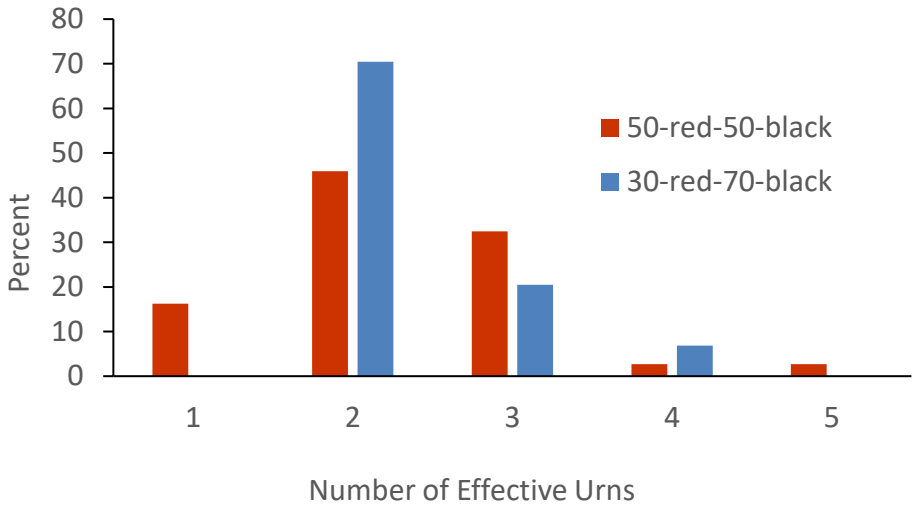


Figure 5: Frequency of Number of Effective Urns

## 5. Concluding Remarks

In this paper, we introduce reciprocation incentives into the Bayesian persuasion model and derive novel implications concerning receiver's responses and sender's optimal information design. Specifically, we find that receivers are persuaded to take the sender's preferred action only if the sender is willing to offer sufficiently informative signals. This result is analogous to the finding in ultimatum game experiments that responders often demand a "fair" division of the surplus and are willing to punish low offer by rejecting it. Understanding this, the proposer typically makes offers that are much more generous than the prediction of subgame-perfect Nash equilibrium. In a similar vein, if the receiver in a persuasion setting has reciprocal preference, the sender needs to design more revealing information structure in order to successfully persuade the receiver. Furthermore, when the prior belief

[6] The theories can only predict the number of "effective" urns because Player A can design two different urns with identical composition (i.e., two different signals with identical likelihood ratios of states).
[7] We group urns with the similar percentage of red balls (within a three-percent difference) as one effective urn.

of the state is more favorable, the sender is deemed to have a higher (ex-ante) expected payoff, so the receiver would be more demanding in the sender's information revelation. Results of laboratory experiments we conduct support these predictions.

There are a number of exciting avenues for future research. First, the effect of reciprocity concern in other communication settings, such as cheap talk and disclosure of verifiable information, can be considered. It is interesting to investigate, theoretically and possibly experimentally, whether reciprocity concern improves or worsens the quality of information transmission in these settings. Second, the simplicity of our experiment design makes it possible to investigate persuasion behaviors in settings with multiple senders or multiple receivers.[8] Finally, as receivers are likely to suffer bias in information processing (such as confirmatory bias and bounded memory), it is interesting to study how the optimal dynamic persuasion technique may exploit these biases.[9]

[8] For multiple receivers, Alonso and Câmara (2016) consider persuading multiple voters in an election. For multiple senders, Au and Kawai (2017b) consider multiple competing senders persuading a receiver to sponsor their own proposals.

[9] For theoretical investigations on dynamic persuasion, see for example, Au (2015) and Ely et.al. (2015).

## References


Alonso, R. and Câmara, O., 2016. Persuading voters. American Economic Review, 106(11), pp.3590-3605.

Au, P.H., 2015. Dynamic Information Disclosure. RAND Journal of Economics, 46(4), pp. 791-823.

Au, P.H., and Kawai, K., 2017a. Competitive disclosure of correlated information. Working paper.

Au, P.H., and Kawai, K., 2017b. Competitive information disclosure by multiple senders. Working paper

Berg, J., Dickhaut, J., and McCabe, K., 1995. Trust, reciprocity, and social history. Games and Economic Behavior, 10(1), 122-142.

Charness, G., 2004. Attribution and reciprocity in an experimental labor market. Journal of Labor Economics, 22(3), pp.665-688.

Charness, G. and Haruvy, E., 2002. Altruism, equity, and reciprocity in a gift-exchange experiment: an encompassing approach. Games and Economic Behavior, 40(2), pp.203-231.

Charness, G. and Levine, D.I., 2007. Intention and stochastic outcomes: An experimental study. Economic Journal, 117(522), pp.1051-1072.

Charness, G. and Rabin, M., 2002. Understanding social preferences with simple tests. Quarterly Journal of Economics, 117(3), pp.817-869.

Cialdini, R.B., 2006. Influence: The psychology of persuasion. Harper Business.

Dufwenberg, M. and Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games and Economic Behavior, 47(2), pp.268-298.

Ely, J., Frankel, A. and Kamenica, E., 2015. Suspense and surprise. Journal of Political Economy, 123(1), pp.215-260.

Falk, A., Fehr, E., and Fischbacher, U., 2003. On the nature of fair behavior. Economic Inquiry, 41(1), 20-26.

Falk, A. and Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior, 54(2), pp.293-315.

Fehr, E., Kirchsteiger, G. and Riedl, A., 1993. Does fairness prevent market clearing? An experimental investigation. Quarterly Journal of Economics, 108(2), pp.437-459.

Fehr, E. and Schmidt, K. M., 1999. A theory of fairness, competition, and cooperation. Quarterly Journal of Economics, 114(3), pp.817-868.

Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics, 10(2), 171-178.

Frechette, G., Lizzeri, A., and Perego, J., 2017. Rules and commitment in communication. Working paper.

Geanakoplos, J., Pearce, D. and Stacchetti, E., 1989. Psychological games and sequential rationality. Games and Economic Behavior, 1(1), pp.60-79.

Gehlbach, S. and Sonin, K., 2014. Government control of the media. Journal of Public Economics, 118, pp.163-171.

Goldstein, I, and Leitner, Y., 2015. Stress tests and information disclosure. Working paper.

Jehiel, P., 2014. On transparency in organizations. Review of Economic Studies, 82(2), pp.736-761.

Güth, W. and Kocher, M.G., 2014. More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature. Journal of Economic Behavior and Organization, 108, pp.396-409.

Güth, W., Schmittberger, R., and Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization, 3(4), 367-388.

Kagel, J.H., Kim, C. and Moser, D., 1996. Fairness in ultimatum games with asymmetric information and asymmetric payoffs. Games and Economic Behavior, 13(1), pp.100-110.

Kamenica, E., and Gentzkow, M., 2011. Bayesian persuasion. American Economic Review, 101(6), 2590-2615.

Nguyen, Q., 2017. Bayesian persuasion: evidence from the laboratory. Working paper.

Rabin, M., 1993. Incorporating fairness into game theory and economics. American Economic Review, pp.1281-1302.

Rayo, L., and Segal, I., 2010. Optimal information disclosure. Journal of Political Economy. 118(5): 949-987.

Roth, A., 1995. Bargaining experiments. In: Kagel, J., Roth, A. (Eds.), Handbook of Experimental Economics. Princeton Univ. Press, Princeton, pp. 253–348.

Schweizer, N. and Szech, N., 2018. Optimal revelation of life-changing information. Management Science.

Wu, J., 2018. Benefits from non-competing persuaders. Working paper.

## Appendix

*Proof of Proposition 1* Let $(p^*, q^*)$ be the sender's equilibrium strategy. Consider the receiver's problem after observing an urn with a fraction $p \geq \mu$ of red balls. His objective function is linear in σ. It is therefore clear that the optimal value of σ is 1 if the coefficient in front of σ is positive; 0 if it is negative; and any value in [0,1] is optimal if it is zero. Moreover, the coefficient in front of σ is positive if and only if

$$\sigma_R(p) < \frac{\frac{(2p-1)(p-q^*)}{\lambda_R} + (1-\mu)(p-q^*) - 2(p-\mu)(1-q^*)\sigma_R(q^*)}{2(\mu-q^*)(1-p)} \equiv D(p).$$

Consequently, for $p \in [\mu, 1]$, $\sigma_R(p)$ is uniquely given by

$$\sigma_R(p) = \begin{cases} 0, \ if\ D(p) < 0 \\ D(p), if\ D(p) \in [0,1]. \\ 1, if\ D(p) > 1 \end{cases}$$

Following similar computation, if we define

$$D(q) \equiv \frac{\frac{(2q-1)(p^*-q)}{\lambda_R} + (1-\mu)(p^*-q) - 2(\mu-q)(1-p^*)\sigma_R(p^*)}{2(p^*-\mu)(1-q)},$$

then $\sigma_R(q)$, for $q \in [0, \mu]$, is uniquely given by

$$\sigma_R(q) = \begin{cases} 0, if\ D(q) < 0 \\ D(q), if\ D(q) \in [0,1]. \\ 1, if\ D(q) > 1 \end{cases}$$

Next, observe that

$$\frac{\partial}{\partial q} D(q) = \frac{2q^2 - 4q + p^* + 1 + \lambda_R(1-p^*)(1-\mu)}{2\lambda_R(p^*-\mu)(1-q)^2}$$

$$\geq \frac{(1-2q)(1-q) + \lambda_R(1-p^*)(1-\mu)}{2\lambda_R(p^*-\mu)(1-q)^2} \geq 0,$$

as $q \leq \mu \leq \frac{1}{2}$ in our model. Therefore, $D(q) \leq D(\mu) \leq \frac{1}{2}$ and $\sigma_R(q) \leq \frac{1}{2}$. Similarly,

$$\frac{\partial}{\partial p} D(p) = \frac{-2p^2 + 4p - q^* - 1 + \lambda_R(1 - 2\sigma_R(q^*))(1-q^*)(1-\mu)}{2\lambda_R(\mu-q^*)(1-p)^2}.$$

As $\sigma_R(q^*) \leq D(\mu)$, the numerator is no less than $2(2p - p^2 - \mu - q^*(1-\mu))$, which in turn is no less than $2(p-\mu)(2-\mu-p)$ as $q^* \leq \mu$. Therefore, $D(p)$ and hence $\sigma_R(p)$ is weakly increasing in $p$. We have thus established that the function $\sigma_R(\cdot)$ is weakly increasing and strictly so whenever $\sigma_R(\cdot) \in (0,1)$.

Furthermore,

$$\frac{\partial^2}{\partial p^2} D(p) = \frac{(1-q)\big(1-\lambda_R(2\sigma_R(q)-1)(1-\mu)\big)}{\lambda_R(\mu-q)(1-p)^3} > 0.$$

As $\sigma_R(.)$ is strictly convex for $p \geq \mu$, it is suboptimal for the sender to choose a $p^*$ such that $\sigma_R(p^*) < 1$. We thus have that $\sigma_R(p^*) = 1$.

To show that $q^* = 0$, it suffices to show that given $\sigma_R(p^*) = 1$, we have $\frac{D(q)}{q} < \frac{1}{p^*}$ for all $q \in [0, \mu]$. Using the definition of $D(q)$ and that assumption that $\lambda_R \leq 1$, the last inequality holds if and only if

$$(2q-1)(p^*-q)p^* + (1-\mu)(p^*-q)p^* - 2p^*(\mu-q)(1-p^*) - 2q(p^*-\mu)(1-q) < 0.$$

It follows from straightforward computation that the left-hand side of the inequality above is strictly increasing in $q$. As $q \leq \mu$ by definition, the inequality above holds if $\mu(p^*-\mu)(p^*+2\mu-2) \leq 0$, which is true as $\mu \leq 0.5$.

Finally, substituting $q^* = 0$ into $D(p^*) = 1$ gives the equilibrium value of $p^*$. Moreover, substituting $\sigma_R(p^*) = 1$ and $q^* = \sigma_R(q^*) = 0$ into the functions $D(p)$ and $D(q)$ defined above gives the receiver's equilibrium strategy. Q.E.D.

*Proof of Corollary 2* In equilibrium, the sender's expected monetary equilibrium payoff is given by $\frac{\mu}{p^*}$. Using Proposition 1, and taking derivative with respect to $\mu$, we have

$$\frac{\partial}{\partial \mu}\left(\frac{\mu}{p^*}\right) = 4\frac{(1-\lambda_R)\sqrt{(1-\lambda_R(1+\mu))^2+16\mu\lambda_R} + \mu\lambda_R(7+\lambda_R) + (1+\lambda_R)^2}{\sqrt{\big(1-\lambda_R(1+\mu)\big)^2+16\mu\lambda_R}\left((1-\lambda_R(1+\mu)+\sqrt{\big(1-\lambda_R(1+\mu)\big)^2+16\mu\lambda_R}\right)^2}$$

It is clear that the derivative is positive.

Next consider the receiver's expected equilibrium monetary payoff, which is given by $\frac{\mu}{p^*} \times p^* + \left(1-\frac{\mu}{p^*}\right) \times 1 = 1-\mu(\frac{1}{p^*}-1)$. Using Proposition 1, and taking derivative with respect to $\mu$, we have

$$\begin{aligned}&\frac{\partial}{\partial \mu}\left(1-\mu(\frac{1}{p^*}-1)\right)\\ &= 1-4\frac{(1-\lambda_R)\sqrt{(1-\lambda_R(1+\mu))^2+16\mu\lambda_R} + \mu\lambda_R(7+\lambda_R) + (1+\lambda_R)^2}{\sqrt{\big(1-\lambda_R(1+\mu)\big)^2+16\mu\lambda_R}\left((1-\lambda_R(1+\mu)+\sqrt{\big(1-\lambda_R(1+\mu)\big)^2+16\mu\lambda_R}\right)^2}\end{aligned}$$

Straightforward algebra shows that the derivative is positive if and only if ${\lambda_R}^2(1+\mu)^2 + (14\mu-4)\lambda_R - 5 > 0$, or equivalently, $\lambda_R > \frac{(2-7\mu)+3\sqrt{1-2\mu+6\mu^2}}{(1+\mu)^2}$. Q.E.D.

## Appendix (Experimental Instructions (50-red-50-black treatment); for online publication)

**Instructions**

Welcome to our experimental study on decision-making. You will receive a show-up fee of HKD40. In addition, you can gain more money as result of your decisions in the experiment.

You will be given a subject ID number. Please keep it confidentially. Your decisions will be anonymous and kept confidential. Thus, other participants won't be able to link your decisions with your identity. You will be paid in private, using your subject ID, and in cash at the end of the experiment.

When you have any questions, please feel free to ask by raising your hand, one of our assistants will come to answer your questions. Please DO NOT communicate with any other participants.

---

**General Instructions**

In this experiment, in each round, participants will be randomly matched into pairs. In each matched pair, one participant is assigned to the role of Player A; and the other the role of Player B. Your role will remain fixed throughout this stage of the experiment. You will not be told the identity of the participant you are matched with, nor will that participant be told your identity – even after the end of the experiment.

This game consists of 10 rounds of decision-making. One round will be randomly selected at the end of the experiment for payment. The game has two steps, which we will describe as follows.

**Step 1**

In each round, there are 50 red balls and 50 black balls. Player A needs to allocate all of the 100 balls into 5 urns. Player A will fill in the following table:

| Urn | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Red | | | | | |
| Black | | | | | |

Player A will determine the number of red balls and black balls in each of the urns. An empty urn is allowed. All of the 100 balls must be allocated to one of the urns; in other words, the sum of balls in the urns must be 100.

Then, one urn will be randomly drawn. The total number of balls in each urn will determine the probability that the urn will be drawn. For example, if urn 1 contains 80 balls, then this urn will be drawn will probability 0.8.

Player A can check the total number of balls allocated "Check" at the bottom of the table. If Player A has finished filling in the table, he/she can click "OK" to submit his/her decision. No further changes to the table can be made after clicking "OK".

**Step 2**

After an urn is drawn, Player B will see a message on the screen specifying the number of red balls and number of black balls in the urn. For example:

*The urn contains 20 red balls and 20 black balls.*

Then, a ball will be randomly drawn from the urn, and player B will guess whether the ball drawn is red or black.

**Payoff**

Player A gets HKD 40 if player B guesses that the ball is red, regardless of the actual color of the ball drawn. Player A gets HKD 0 if player B guesses that the ball is black, regardless of the actual color of the ball drawn.

If player B's guess is the same as the ball drawn, he/she gets HKD 40. If the guess is incorrect, his/her payoff is zero.

**Information Feedback**

At the end of each round, you will be informed about (i) the urn drawn, (ii) the guess made by Player B, (iii) the color of the ball drawn, and (iv) your earning.

**Practice Round**

We will provide you with one practice round. At the beginning of the practice round, you will be randomly assigned the role of either Player A or Player B. Your role in the official rounds will be the same as that in the practice round. Once the practice round is over, the computer will tell you "The official rounds begin now!"

**Administration**

Your decisions and your monetary payment will be kept confidential. Upon finishing the experiment, you will receive your cash payment. You will be asked to sign your name to acknowledge your receipt of the payment. You are then free to leave.

If you have any question, please raise your hand now. We will answer your question individually. If there is no question, we will proceed to the practice round now.

**Subject ID:**

**Questionnaire**

Please answer Q1 if you are player A.

Please answer Q2 if you are player B.

Q1. Please explain how you determine the allocation of balls in the urns.

Q2. Please explain how you choose between guessing on red and black ball.