# Incentives or Persuasion? An Experimental Investigation*

**Andreas Aristidou**[†] **Giorgio Coricelli**[‡§] **Alexander Vostroknutov**[¶]

March 2023

**Abstract**

A principal can influence agent's actions through monetary transfers (Mechanism Design) or Bayesian persuasion (Information Design). We provide an experimental comparison of these incentive structures using theoretically equivalent games. Behavior in Information Design is in line with the theory, however the equivalence breaks down in Mechanism Design due to agents' high monetary demands. Principals extract higher rents by persuading than incentivizing agents. Nevertheless, agents are able to capture a significant proportion of principals' surplus due to their bargaining power arising from outside options. While equivalent in theory, Mechanism and Information Design produce different behavior and welfare effects.

JEL classifications: *C91, C92, D47, D91.*
Keywords: *Bayesian persuasion, mechanism design, information design, experiments.*

*For helpful discussions, in alphabetical order, we thank Ali Abboud, Raphael Boleslavsky, Odilon Câmara, Gabriele Camera, Ambuj Dewan, Laura Doval, Ignacio Esponda, Guillaume Frechétte, Cary Frydman, Simone Galperti, Matt Gentzkow, John Hatfield, Agne Kajackaite, Ian Krajbich, Yilmaz Kocer, Anton Kolotilin, Philippos Louis, Tomaz Sadzik, Joel Sobel, Omer Tamuz, Severine Toussaert, Nikolaos Tsakas, Joel Watson. This paper benefited greatly from discussions at the 5th workshop on behavioral game theory at the University of East Anglia, the 2019 North American ESA conference at Loyola Marymount and the University of Southern California, the 2nd annual Los Angeles Conference in Applied Economics, the 93rd annual conference of the Wester Economic Association International, and the brown-bag seminars at the Department of Economics at USC. GC acknowledges the financial support of the European Research Council (ERC Consolidator Grant 617629). All errors are our own.

†Department of Economics, University of Southern California, 3620 South Vermont Ave. 300, Los Angeles, CA 90089, USA. e-mail: aristidouandreas13@gmail.com.

‡Department of Economics, University of Southern California, 3620 South Vermont Ave. 300, Los Angeles, CA 90089, USA. e-mail: coricell@usc.edu.

§Corresponding author.

¶Department of Microeconomics and Public Economics, Maastricht University, Tongersestraat 52, 6211LM Maastricht, The Netherlands. e-mail: a.vostroknutov@maastrichtuniversity.nl.

# 1 Introduction

There are two types of strategies that a rational principal can employ to influence agent's behavior. The first strategy is to use monetary incentives to alter the size of the potential payoffs that the agent can expect from each available action, and the second is to use Bayesian persuasion to alter the likelihood of the potential payoffs resulting from each action. For example, consider an online retailer (the principal) who attempts to convince a consumer (the agent) to buy some product. The consumer faces uncertainty over the quality of the product and prefers to buy if the quality is high and not to buy if the quality is low. The consumer also holds a prior belief about the likelihood of the quality being high and acts to maximize his expected utility. What can such online retailers do to increase the chances that their potential consumers find it optimal to purchase? First, they can provide monetary incentives—like discounts, promotions, or bundling—directed at reducing consumers' costs. Second, they can utilize Bayesian persuasion, for example, design personalized recommendations, expert reviews, or product placement, in order to shift the consumers' prior beliefs about the quality of the product.

While traditionally the design of mechanisms with monetary transfers was used in the majority of studies on principal-agent problems (Mechanism Design), more recently, the literature has shifted focus to the design of strategic information disclosure (à la Kamenica and Gentzkow, 2011) as a robust way of influencing economic agents (Information Design). The timing of these recent developments in strategic information disclosure is no coincidence. The beginning of the informational era of big data and machine learning has given way to massive information gatekeepers (e.g., Yelp, Netflix, Amazon, etc.) who collect and utilize information to guide consumers' behavior. This has been gradually shifting the focus of interest from monetary incentives to Bayesian persuasion.[1]

The goal of our study is to conduct a first *comparative analysis*, both theoretical and experimental, of the two strategies that principals can use to influence agents' choices. Thus, we try to equate the influence of various factors across the two environments to have as clean a comparison as possible between persuasion with *monetary* and *informational* incentives. To this end, we construct two games that model Mechanism and Information Design interactions in a way that makes them theoretically equivalent. We conduct an experiment to compare the behavior of participants in these games and try to shed some light on whether the games' theoretical equivalence holds in practice; and if not, what is the difference in actual choices and welfare of participants from the theoretical predictions. We believe that this analysis can provide important insights into the practical issues that may arise when Mechanism or Information Design are used in real-life applications.

We draw our inspiration from a recent literature that attempts to study Bayesian persuasion in a general game-theoretic framework, making parallels with mechanism design and naming it

[1]Throughout the paper we use the terms *monetary incentives / Mechanism Design;* and *informational incentives / Bayesian persuasion / Information Design* interchangeably.

Information Design.[2] The studies by Bergemann and Morris (2017) and Taneva (2017) examine the problem of a designer who seeks to impose an agenda on a group of players. As is conventional in mechanism design, the designer is assumed to have the ability to commit to a transfer mapping to the players. However, in the parallel world of information design, the information designer is instead assumed to have an informational advantage over the players by being able to commit to a signal structure (probabilistic state-message mapping), essentially recommending actions to players.[3]

This theoretical parallelism is both intriguing and exciting for game theorists and economists in general. The mechanism design problems explored in many original studies of the past decades can now be investigated through an alternative route, namely information design. From an applied perspective however, this raises several natural questions. How does this theoretical parallelism play out in practice? Can information designers utilize persuasion with the same effect (i.e., to maximize their payoffs) as mechanism designers use monetary incentives? How do people react to being persuaded rather than incentivized? These questions define the scope of our paper. In a simple bilateral setting, where principals can act as both information and mechanism designers, we test whether they are more successful in using incentives or persuasion to influence the agents' choices and to increase their payoffs.

In the theoretical part of the paper we employ the original two-state Bayesian persuasion model by Kamenica and Gentzkow (2011). We start with a baseline setup in which a principal is not able to act as either an information or a mechanism designer and trivially show that, in this case, she is guaranteed to get the lowest possible payoff. Then we extend the baseline game in two directions: 1) the principal can act as an information designer in an attempt to persuade the agent to take the principal-preferred action by communicating informative recommendations and 2) the principal can act as a mechanism designer in an attempt to incentivize the agent to take the principal-preferred action by providing monetary transfers. We show that the two games are equivalent in terms of best response correspondences and most importantly that the expected payoffs of both players are identical in what Kamenica and Gentzkow (2011) call the "Principal-Preferred" Subgame Perfect Equilibrium [(PP)SPE].[4]

With this theoretical equivalence in place, we have a foundation on which we can test our research questions. Experimentally, however, we face two problems: 1) Bayesian persuasion in its standard form is too computationally intensive for lab participants and 2) Information Design (ID)

[2]The literature on information design is expanding fast with many recent contributions (Alonso and Câmara, 2016a,b,c; Babichenko and Barman, 2016; Bergemann and Morris, 2016; Bizzotto *et al.*, 2016; Boleslavsky and Kim, 2018; DellaVigna and Gentzkow, 2010; Dughmi and Xu, 2016; Dughmi *et al.*, 2016; Dughmi and Xu, 2017; Gentzkow and Kamenica, 2014, 2016, 2017; Gratton *et al.*, 2017; Hernández and Neeman, 2018; Kolotilin *et al.*, 2017; Li and Norman, 2018; Wang, 2013). The early seminal papers include Crawford and Sobel (1982) and Okuno-Fujiwara *et al.* (1990).

[3]For another interesting approach see Mathevet *et al.* (2020).

[4]Actually, Kamenica and Gentzkow (2011) name it "Sender-Preferred" Subgame Perfect Equilibrium to emphasize that in all of the cases where the receiver is indifferent between actions she always chooses the one that the sender prefers. We do the same except that, for reasons pertaining to comparison with the mechanism design literature, we call the Sender *Principal* and the Receiver *Agent*.

and Mechanism Design (MD) are very different games, thus in order to detect behavioral differences that arise exclusively from their inherent features (information vs. money), the experimental setup should maximize the similarity of the two choice environments and eliminate any other confounds. To solve these problems, we propose an innovative experimental design, which not only minimizes the differences between the two games, but also renders Bayesian persuasion a relatively "user-friendly" task. In our design, ID and MD tasks are simplified so that subjects in the roles of both principals and agents choose a single number.

We find that principals, who attempt to influence the agents to choose their preferred action, extract higher rents when using informative recommendations in ID than when using monetary incentives in MD, despite the equivalence predicted theoretically. Specifically, principals are able to persuade agents more often than they are able to incentivize them, and successful persuasion attempts are, on average, more profitable than successful incentivization attempts. This result seems to be at least partially driven by the fact that agents appear to be more demanding when they are being incentivized than when they are being persuaded.

Similar to Fréchette *et al.* (2022), the average behavior in ID is in line with the theoretical predictions in all periods. However, in MD we observe a different pattern: with time, principals adjust their demands to those of the agents, which leads to principals' payoffs being lower than theory predicts. Analyzing the percentage of successful transactions in time, we conclude that this pattern emerges because principals try to avoid zero payoff that they get when they fail to persuade or incentivize the agent.

This analysis suggests that principals' low average payoffs in both games (as compared to the equilibrium) can be at least partially attributed to the distribution of *bargaining power* that is ignored in theoretical models of both Mechanism and Information Design. While the standard theoretical analysis of the two-stage non-cooperative game played by our participants gives zero bargaining power to the agents, we find that they are able to seize some part of the surplus, as if they have 40% of the bargaining power (as predicted by Nash Bargaining Solution), a result that is remarkably similar in both ID and MD environments. We believe that this is the case since principals have can incur a loss from unsuccessful attempt at persuasion or incentivization – they receive the payoff of 0, whereas agents can guarantee themselves a decent payoff no matter what principal does. This feature of principal-agent problems is not specific to our experiment and thus can have tangible implications in practice.

Overall, our study shows that, while equivalent in theory, monetary and informational incentives can have dramatically different effects on behavior in reality. Our experiment identifies several dimensions where behavior is differentially affected by the use of monetary versus informational incentives. We also believe that the insights from our analysis can motivate the development of more behaviorally-driven theories of persuasion and incentivization that could help to better understand the applied side of the theoretical parallelism between the two incentive structures.

# 2 Theoretical Framework

The results derived from the theoretical framework presented in this section serve as the motivation behind our research questions and the experimental design. Our framework is a version of the original two-state Bayesian persuasion setup of Kamenica and Gentzkow (2011). We begin with an adaptation of the model where the principal (sender) is unable to send messages to the agent (receiver), thus making him a mere observer. This results in the expected utility maximizing agent (receiver) choosing the action that gives the principal the smallest payoff. We then extend this baseline setup in two independent directions. In the first extension, the principal can commit to probabilistic state-contingent messages to the agent (persuasion) in the same way as in Kamenica and Gentzkow (2011). We call this the "information design extension." In the second extension, the principal can instead commit to action-contingent transfers to the agent (monetary incentives). We call this the "mechanism design extension." Finally, the agent observes the principal's message or incentives and takes an action. Both games admit a unique Subgame Perfect Equilibrium in which the expected payoffs of principals/agents are identical in the two games. More precisely, while the principal benefits from having the ability to commit to messages (persuasion) or transfers (incentives), she is indifferent between the two. The agent neither gains nor loses from the principal's ability to persuade or incentivize.

## 2.1 The Baseline Model

Suppose that there are two players Principal ($P$) and Agent ($A$), and two states of the world $S = \{R, B\}$ (red or blue ball) happening with probabilities $\Pr(B) > \Pr(R) \equiv p$, which is common knowledge. The agent's action set is $C^A = \{r, b\}$, and the Principal's action set is empty, $C^P = \{\emptyset\}$. The state-action contingent payoffs for each player are denoted by $\Pi^i_{s,c} \in \mathbb{R}$, where $i \in \{A, P\}$ refers to the player, $s \in S$ refers to the realized state and $c \in C^A$ refers to the agent's action. The agent receives positive payoff if she chooses the action which matches the state (i.e., $c = r$ when $s = R$ or $c = b$ when $s = B$). The principal's payoffs are state-independent. She is solely interested in the agent's action and receives positive payoff only when the agent takes action $r$. We assume that the payoffs adhere to the following restrictions: $\Pi^A_{R,r} = \Pi^A_{B,b} \equiv \Pi^A > 0$, $\Pi^A_{R,b} = \Pi^A_{B,r} = 0$, $\Pi^P_{R,r} = \Pi^P_{B,r} \equiv \Pi^P > 0$, $\Pi^P_{R,b} = \Pi^P_{B,b} = 0$. To achieve the equality of the expected payoffs in equilibrium in the information design and mechanism design extensions, we need to impose $\Pi^A = \Pi^P \equiv \Pi$. The payoffs are summarized in Table 1.

| | | State realization | |
|---|---|---|---|
| | | $R$ | $B$ |
| Agent's choice | $r$ | $\Pi$ , $\Pi$ | $\Pi$ , 0 |
| | $b$ | 0 , 0 | 0 , $\Pi$ |

Table 1: Payoff matrix in the baseline model. In each cell, the leftmost number represents the principal's payoff.

The agent maximizes her expected payoff and thus will always choose $b$, which maximizes the (ex-ante) probability of matching the state. Given the agent's optimal action $c^* = b$, the expected utilities of the players are

$$E_s\Pi^P_{s,c^*} = 0 \qquad \text{(Principal}^{\text{Baseline}}\text{)}$$
$$E_s\Pi^A_{s,c^*} = (1-p)\Pi. \qquad \text{(Agent}^{\text{Baseline}}\text{)}$$

## 2.2 Information Design Extension

Consider the following extension of the baseline model where the principal can act as an information designer (Stage 1) prior to the agent's choice (Stage 2). The principal is now endowed with the ability to construct a state-message mapping (henceforth, a "signal structure") from which a message $m$—correlated with the realized state of the world—is communicated to the agent. This mapping determines the level of correlation between the states of the world and the messages (or the informativeness of each message). The principal's action set becomes $C^P = \{(P_R, P_B) \mid P_R, P_B \in [0,1]\}$, where $P_R = \Pr(m = \rho | s = R), P_B = \Pr(m = \beta | s = B)\}$ are the probabilities of a message $m \in M = \{\rho, \beta\}$ that is to be communicated to the agent [5] Knowing $(P_R, P_B)$ and $m$, the agent Bayes-updates her beliefs about the likelihood of each state based on the message received and the signal structure from which the message was generated. Given her updated beliefs, the agent maximizes her expected payoff by choosing action $c^*$, which matches the state that is more likely to have been realized. Implicit in this are two critical assumptions: 1) the principal is able to condition the messages on the realized state of the world without having observed it and 2) the principal can credibly commit to the signal structure (i.e., the agent can observe the mapping which generated the message). Without loss of generality, we assume $|M| = |S|$ (see Kamenica and Gentzkow, 2011). In this way, messages can be thought of as action recommendations. The unique Principal-Preferred Subgame Perfect Equilibrium [(PP)SPE] of this game admits the following expected payoffs:[6]

$$E_s\Pi^P_{s,c^*} = 2p\Pi \qquad \text{(Principal}^{\text{ID}}\text{)}$$
$$E_s\Pi^A_{s,c^*} = (1-p)\Pi \qquad \text{(Agent}^{\text{ID}}\text{)}$$

Note the following: 1) the principal uses Bayesian persuasion (information design) to increase her expected payoff from zero (baseline model) to $2p\Pi$, by providing state-contingent messages (action recommendations) that the agent finds optimal to follow rather than ignore; 2) the agent neither benefits nor loses from the principal's persuasion. The latter happens because, while the principal tries to extract as much surplus as possible, she is constrained by the expected payoff that the agent

[5] A message $\rho$ or $\beta$ is always sent. When the ball is red ($s = R$), the message $\beta$ is sent with probability $1 - P_R$ and when the ball is blue ($s = B$) the message $\rho$ is sent with probability $1 - P_B$.

[6] See Appendix A.1 for derivations.

can guarantee herself by simply choosing $b$, which we will refer to as the agent's outside option. Thus, all social surplus generated by information design is captured by the principal.

## 2.3 Mechanism Design Extension

Consider another extension of the baseline model, where the principal can act as a mechanism designer (Stage 1) prior to the agent's decision (Stage 2). The principal is now able to construct an action-contingent mapping which determines how payoffs are to be transferred from the principal to the agent conditional on the agent's action. As a mechanism designer, the principal's action set becomes $C^P = \{(t_r, t_b) \mid t_r, t_b \in [0, \Pi]\}$, where $t_r$ and $t_b$ are the transfers to the agent if she chooses action $r$ or $b$ respectively. The agent takes into account the additional conditional payoffs and chooses action $c^*$ that maximizes her overall expected payoff.[7] Implicit in this are two assumptions: 1) the principal is able to condition the transfers on the agent's actions; 2) the principal can credibly commit to the action-contingent transfers (i.e., the agent is guaranteed to receive the transfer that is contingent on her chosen action). The unique Principal-Preferred Subgame Perfect Equilibrium of this game admits the following expected payoffs:[8]

$$E_s\Pi^P_{s,c^*} = 2p\Pi \qquad \text{(Principal}^{\text{MD}}\text{)}$$
$$E_s\Pi^A_{s,c^*} = (1-p)\Pi \qquad \text{(Agent}^{\text{MD}}\text{)}$$

Note that 1) the principal uses monetary incentives (mechanism design) to increase her expected payoff from 0 (baseline model) to $2p\Pi$, by providing action-contingent transfers which induce the agent to choose action $r$ instead of her baseline-optimal action $b$; 2) the agent neither benefits nor loses from the principal's incentivization. The agent's expected payoff remains the same as in the baseline game, and thus all social surplus from mechanism design is captured by the principal.

We summarize the equilibrium expected payoffs from the baseline model and the two extensions:

$$\begin{array}{rclcl} \text{Principal}^{\text{Baseline}} & < & \text{Principal}^{\text{ID}} & = & \text{Principal}^{\text{MD}} \\ \text{Agent}^{\text{Baseline}} & = & \text{Agent}^{\text{ID}} & = & \text{Agent}^{\text{MD}} \end{array}$$

The equality of the principal's equilibrium expected payoffs in the information design and the mechanism design extensions of the baseline model is our theoretical object of interest that the experiment described below is designed to test.

[7]Here "additional" refers to the conditional payoffs that the agent receives in addition to the payoffs described in Table 1.

[8]See Appendix A.2 for derivations.

# 3 Experiment

The experiment consisted of two treatments with two sections in each: Section ID (information design) and Section MD (mechanism design). The two treatments differed in the order of sections ID and MD. This allowed us to investigate possible order effects in the ID and MD sections. At the beginning of the experiment participants were randomly assigned one of two possible roles: Principal (denoted as "Player A") or Agent (denoted as "Player B"), the roles that were fixed throughout the experiment. In Section ID participants played 10 periods of the information design game, and in Section MD they played 10 periods of the mechanism design game. In each period, every principal was randomly matched with one agent to form a pair and to play the respective game. At the end of each period each participant received feedback about the outcome of the game and points earned. Participants were paid for one randomly chosen period from each section with an exchange rate of 100 points corresponding to 5 Euros (thus, they were paid for two choices in total). Experimental instructions and screenshots can be found in Appendix C.

In each period and for every pair, participants were told that a ball would be randomly drawn from a virtual urn with 10 balls, three of which were red and seven blue. The goal of the agent was to correctly guess the color of the ball, and the goal of the principal was to attempt to persuade (section ID) or incentivize (section MD) the agent to guess red. The color of the ball was revealed to the participants at the end of the period. Experimental points earned by each participant depended on the revealed color and the agent's guess in accordance with the games in our models with $p = 0.3$ and $\Pi = 100$.

While the ID and MD games in our theoretical framework are described as two-stage sequential games, we implemented both as simultaneous-move games. The principal constructed a signal structure (ID) or transfers (MD) and, at the same time, the agent chose which signal structures she wished to follow (ID) or which transfers to accept (MD). We describe this in more detail below.

## 3.1 Section ID (Information Design)

**Principal's choice**. The principal's role in ID was to act as an information designer in accordance with the ID game described in Section 2.2. Specifically, the principal had to construct a signal structure ($P_R$, $P_B$) that would generate the recommendation "Guess red" or "Guess blue" conditional on the color of the ball drawn.[9] To simplify the decision problem and following the findings of Fréchette *et al.* (2022), we fixed $P_R$ at the equilibrium level of $P_R = 1$ (see Appendix A.1).[10] Consequently, the principal's only choice was to set the percentage chance of generating the correct recommendation when the ball drawn was blue ($P_B$). We denote this choice by $X \in [0\%, 100\%]$.

[9] The signal structure ($P_R$, $P_B$) sets the probabilities with which each recommendation is generated in each state (ball color) as follows: $P_R = \Pr(m = \text{"Guess red"} \mid s = \text{Red ball})$ and $P_B = \Pr(m = \text{"Guess blue"} \mid s = \text{Blue ball})$.

[10] Fréchette *et al.* (2022) allow subjects to manipulate the equivalent of $P_R$ and find that the vast majority of choices are at equilibrium ($P_R = 1$).

Thus, the principal's signal structure is $(P_R, P_B) = (1, X)$. To understand what different choices of $X$ imply it is worth considering two extreme cases. When $X = 100\%$ the recommendation is always correct and fully reveals the color of the ball (Full Information). When $X = 0\%$ the recommendation is always "Guess red," so no information about the color of the ball is provided since when the ball is red the recommendation is also "Guess red" ($P_R = 1$; No Information). For intermediate values of $X$, the higher the $X$ the more information gets revealed.

**Agent's choice**. The agent's role in ID was to determine whether she would follow or ignore the principal's recommendation for all possible signal structures. Following a recommendation means that the agent guesses the color that the recommendation suggests. Ignoring the recommendation means that the agent guesses blue regardless of the recommendation, which maximizes her payoff given the prior beliefs about the urn composition. Without observing the principal's choice of $X$, the agent had to select which signal structures she wished to follow and which ones to ignore by choosing a cutoff minimum value of $P_B$ denoted by $Y \in [0\%, 100\%]$. The elicitation of a cutoff is appropriate here because the informativeness of the signal structure is monotonic in $X$. For example, the agent's expected payoff from following a recommendation from a signal structure $(1, X)$ is greater than that from following the recommendation from any signal structure $(1, X')$ with $X' < X$. By choosing $Y$ in this manner the agent agreed ex ante to follow any recommendation coming from a signal structure that is at least as informative as $(1, Y)$ and to ignore any recommendation coming from less informative signal structures. Thus, the agent was not explicitly asked for a guess. Instead, if the agent followed the principal's signal (when $X \geq Y$), her guess of the color of the ball was determined by the generated recommendation (red or blue), while in the opposite case ($X < Y$), her guess was always the color blue. Agents who choose high values of $Y$ are hard to persuade since they only follow recommendations from very informative signal structures, while agents who choose low $Y$ are easy to persuade and follow recommendations from a large range of $X$.

**The ID interaction**. The principal faces the following tradeoff. Decreasing $X$ yields a higher chance of the "Guess red" recommendation, but a lower chance that it will be followed by the agent ($X \geq Y$). Thus the principal wants to choose $X$ as low as possible conditional on it being weakly greater than $Y$. The agent is less interested in the principal's choice. As long as she doesn't choose $Y$ too low, she is guaranteed a good expected payoff, either by being persuaded or through her outside option (guess blue).

## 3.2 Section MD (Mechanism Design)

**Principal's choice**. The principal's role in MD was to act as a mechanism designer in accordance with the MD game described in Section 2.3. The principal had to choose action-contingent transfers $(t_r, t_b)$ that would be transferred to the agent depending on her guess. In order to make the ID and MD games similar and since the principal earned zero points when the agent's guess was blue, we fixed $t_b$ at the equilibrium level of $t_b = 0$. Consequently, the principal's only choice was to determine the number of points that would be transferred to the agent if the agent's guess was red ($t_r$). We also denote this choice by $X \in [0, 100]$. Thus, the principal was choosing $(t_r, t_b) = (X, 0)$.

**Agent's choice**. The agent's role in MD was to determine whether she would accept or reject the principal's transfer. If the transfer is accepted, the agent committed to guessing red, while if it is rejected the agents committed to guessing blue. The agent had to choose which transfers she wished to accept and which ones to reject by choosing a cutoff minimum value of $t_r$, once again denoted by $Y \in [0, 100]$. Eliciting a cutoff value is also appropriate here since the agent's expected payoff from accepting the transfer is monotonic in $X$. By choosing $Y$ in this manner, the agent agreed ex ante to accept the transfer (and guess red) if it was at least $Y$ points and reject it (and guess blue) if the transfer was less than $Y$ points. Thus, like in ID, the agent was not explicitly asked for a guess. If the agent accepted the principal's transfer (when $X \geq Y$) her guess was red, whereas if the agent rejected the transfer (when $X < Y$) her guess was blue.

**The MD interaction.** The principal faces the following tradeoff. Decreasing $X$ yields a higher payoff if the agent accepts, but also a lower chance of acceptance ($X \geq Y$). Thus, once again, the principal should aim to choose $X$ as low as possible conditional on it being weakly greater than $Y$. The agent is again less interested in the principal's choice. As long as she doesn't choose $Y$ too low, she is guaranteed a good expected payoff, either by being incentivized or through her outside option (guess blue).

## 3.3 Feedback

At the end of every period, participants received feedback about the outcome of the game. Feedback was designed to reflect the information that participants would have received after playing the two-stage ID or MD game. Both participants learned the color of the ball drawn, the recommendation generated (only in ID), the agent's guess and the points earned. Agents also learned the principal's choice of $X$ (the signal structure or transfer) while principals only learned whether agents followed/accepted (i.e., whether $X \geq Y$) or ignored/rejected their recommendation/transfer (i.e., whether $X < Y$). The reason for this asymmetry in feedback is that neither theory nor practice dictate that the principal should learn the precise amount of recommendation informativeness or transfer amount that would induce the agent to follow the recommendation or accept the transfer.

## 3.4 Summary of the Experimental Procedure

In each period (of both sections ID and MD) a principal and an agent were randomly paired. Each participant simultaneously chose a number from 0 to 100 by sliding a pointer (see Appendix C for the screenshots). Principals' choices were referred to as $X$ and agents' choices as $Y$. If $X \geq Y$ we say that the principal has successfully persuaded/incentivized the agent (or that the players have matched). In this case, the agent follows/accepts the principal's recommendation/transfer. This in turn implies that the agent's guess is determined as follows. In Section ID the agent guesses red if the recommendation is "Guess red" and guesses blue if the recommendation is "Guess blue." In Section MD if $X \geq Y$ the agent guesses red and $X$ points are transferred from the principal to the agent. If $X < Y$ we say that the principal has failed to persuade/incentivize the agent (or that the players have not matched). In this case, the agent ignores/rejects the principal's recommendation/transfer. This implies that the agent guesses blue in both games and that no points are transferred in MD.

### 3.5 Design Implementation

The experiment was conducted at the Department of Economics, University of Trento, Italy. We collected the data from 8 sessions with a total of 108 subjects. In the 4 sessions of Treatment 1, participants played the ID section first, and in the 4 sessions of Treatment 2 they played the MD section first. Each session with 12 or more participants was divided into two groups in which random matching was done independently. Thus, the two groups inside such session never interacted creating two independent observations. We have 6 groups with the total of 50 participants in Treatment 1 and 7 groups with 68 participants in Treatment 2, which constitutes 13 independent observations.

Participants were informed that they will take part in an experiment with two parts and that the instructions for each part will be given to them before each part begins. In order to familiarize the participants with the rules of the game and the interface, in Sections ID and MD they played one round of the game in both roles (with themselves). We did not find any significant order effects and thus we merge the data from the two treatments (irrespective of the order of sections ID and MD). Sessions lasted around 1 hour and 30 minutes and participants were paid on average 12 Euros (8 Euros for principals and 14 Euros for agents), a compensation which is in line with the average payment for similar experiments in Italy.

## 4 Results

### 4.1 Incentives or Persuasion?

A natural way to start our comparative analysis of persuasion versus monetary incentives is looking at the average choices $X$ and $Y$ made by principals and agents in the ID and MD environments. In order to be able to use non-parametric tests we consider averages over choices in each of the 13 independent groups of participants described in Section 3.5.[11] Thus, we operate with 13 independent observations. Table 2 shows average choices in all interactions (pairs) and the corresponding theoretical counterparts partitioned by roles and the type of games. In the ID game, principals' and agents' average choices are very close to the theoretical predictions (signed-rank tests, $p = 0.753$ and $p = .311$ respectively), but are significantly higher than the predictions in the MD game (signed-rank

[11] We find no significant order effects between sections ID and MD and so we merge the data from all relevant sections irrespective of the order.

tests, $p < 0.002$). This means that agents appear to be on average more demanding with incentives than with persuasion: in order to comply with the principal they ask for more points in MD than for the equivalent recommendation informativeness in ID. This can make it harder for principals to influence agents' choices through monetary incentives, thus potentially making persuasion a more successful strategy.

| Section | Measure | Sample | Principals | | Agents | |
|---|---|---|---|---|---|---|
| | | | $X$ | Theory | $Y$ | Theory |
| ID | Choices | All pairs | 56.6 (3.01) | 57.1 | 60.4 (3.41) | 57.1 |
| MD | Choices | All pairs | 50.2 (1.52) | 40 | 56.5 (3.43) | 40 |
| $N$ of observations | | | 13 | | 13 | |

Table 2: Average choices in the ID and MD games in 13 independent groups of participants. Numbers in brackets indicate standard errors. "Theory" columns show the theoretical point predictions based on the (PP)SPE. $X$ and $Y$ columns show the averages over the 13 groups.

To understand whether principals are more successful at persuasion or incentivization notice that their earnings critically depend on two factors: 1) how often a principal is able to successfully persuade/incentivize the agents (an unsuccessful attempt gives her zero points); 2) her choices in the periods when she successfully persuades/incentivizes the agents. Thus, a principal is facing a trade-off: increase $X$ (give a more informative recommendation/transfer more money) to improve the chances of a successful persuasion/incentivization and decrease $X$ (give a higher probability weight to her preferred recommendation/transfer less money) to improve the expected payoff from a potentially successful persuasion/incentivization. We find that in ID principals suceed in their persuasion attempts on average in 5.25 periods out of 10, while in MD the success rate is 4.36 periods. The difference is significant (signed-rank test, $p = .043$). Thus, overall principals are more often successful when persuading rather than when incentivizing agents. Still, successful matches constitute only about half of the total attempts.[12] This represents a stark contrast to the theory where a principal can extrapolate the agent's cutoff through backward induction.

Next, we look at the payoffs that principals and agents receive given a successful attempt. To do that we consider average Matched Expected Payoffs (MEP). These are not the actual payoffs observed by the participants, but rather what they should expect to receive given their choices $X$ and $Y$ and conditional on being matched.[13] We find that this is a better measure of players'

[12] As explained in Section 3.4, a match refers to the situation when $X \geq Y$ (a successful persuasion or incentivization). The case $X < Y$ is referred to as a non-match (an unsuccessful persuasion or incentivization).

[13] In case of a match in ID, Matched Expected Payoffs are $(100 - 0.7X, 30 + 0.7X)$ for principals and agents respectively. In case of a match in MD, Matched Expected Payoffs are $(100 - X, 30 + X)$.

aggregate performance than the realized payoffs since MEP do not contain noise due to the random draws of the ball and thus constitute a more natural comparison to the theoretical predictions.[14]

| Section | Measure | Sample | Principals | | Agents | |
|---|---|---|---|---|---|---|
| | | | MEP | Theory | MEP | Theory |
| ID | Payoffs | Matched pairs | 47.8 (2.73) | 60 | 82.2 (2.73) | 70 |
| MD | Payoffs | Matched pairs | 36.6 (3.41) | 60 | 93.4 (3.41) | 70 |
| $N$ of observations | | | 13 | | 13 | |

Table 3: Average Matched Expected Payoffs (MEP) in ID and MD. Numbers in brackets indicate standard errors.

Table 3 shows average MEP in the 13 independent groups of participants. Successful persuasions yield higher payoffs to the principals than successful incentivizations (signed-rank test, $p = .0058$). MEP are higher in ID than in MD in 11 out of 13 groups. Nevertheless, in both games, principals still earn significantly less than the equilibrium prediction (signed-rank tests, ID: $p = .0037$, MD: $p = 0.0015$). This observation is also reflected in the earnings of the agents, who get significantly more than the theoretical predictions (signed-rank tests, ID: $p = .0037$, MD: $p = 0.0015$).

**Result 1.** *Principals make more money by persuading (ID) than by incentivizing (MD): 1) they successfully persuade agents to follow recommendations more often than they manage to incentivize them to accept transfers and 2) successful persuasion attempts yield on average higher payoff for the principals than successful incentivization attempts. This can be partially attributed to agents demanding higher monetary transfers in MD than the equivalent informativeness of the recommendation in ID.*

## 4.2 Period-by-Period Analysis

In order to understand what drives the aggregate results in the previous section we examine the per-period evolution of average choices in the 13 independent groups shown in Figure 1. We can see that while choices in ID remain relatively stable for both principals and agents, in MD we observe principals starting lower than agents in period 1 and then gradually increasing their choices until they reach agents' average choices, which remain relatively stable throughout the game.

This behavior may indicate the existence of an inherently stable level of agents' choices (different in each game) towards which principals converge. In ID principals seem to achieve that point in period 1 and thus do not move away from it. An interpretation is that participants understand the asymmetry in each player's outside option. When not matched, agents earn 70 points on average,

[14]The results with MEP are almost identical with the results with realized payoffs. Realized payoffs both for the matched pairs and for the whole sample can be found in Table 4, Appendix B.

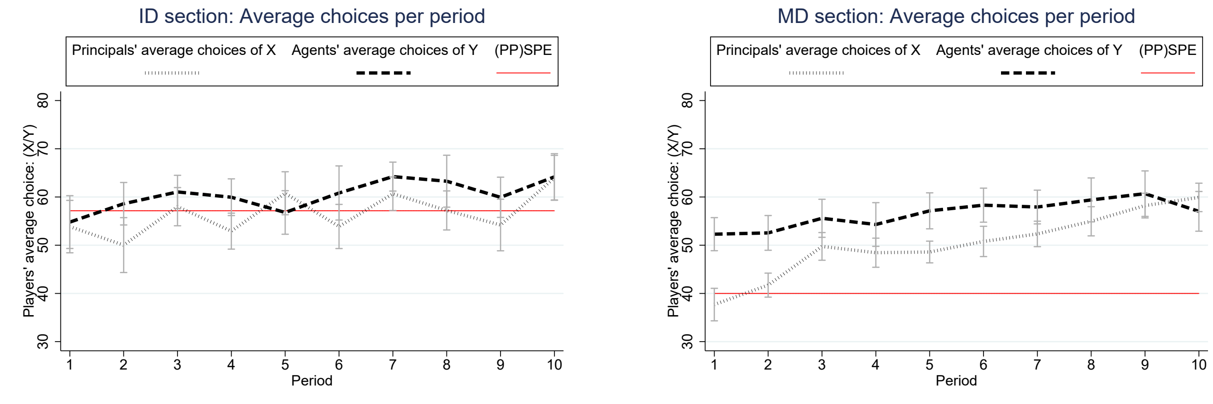



Figure 1: Average choices of principals (dotted lines) and agents (dashed lines) in each period of the ID and MD games. The solid red lines indicate the predictions of the two-stage (PP)SPE. Error bars are ±1SE corresponding to 13 observations.

whereas principals get zero points. Thus, agents remain relatively stable in their choices, while principals are forced to adapt to agents' choices in order to increase their chance of matching.[15] This idea is illustrated by the dynamics of the number of matches displayed on Figure 2. The number of matches in ID is stable, but in MD it grows together with principals' average choices (Figure 1) reaching the levels of ID matches around period 8. It thus seems that a match rate of about 50% is what principals find optimal when resolving the trade-off between successful persuasion/incentivization and recommendation informativeness/transfer amount. This however is not efficient as any unsuccessful persuasion/incentivization attempt results in an expected welfare loss of 60 points.[16] As a consequence, the optimal resolution of this trade-off by the principals brings about a sizeable loss of surplus.

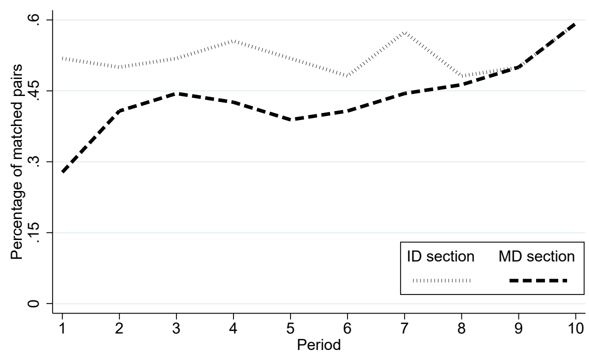



Figure 2: Percentage of matched pairs ($X \geq Y$) per period in each game.

**Result 2.** *The possibility of a non-match creates the situation where principals have to trade-off between choosing low $X$ (high probability of non-match and high profit if matched) and choosing high $X$ (low probability of non-match and low profit if matched). As a result, principals are forced to adjust their offers $X$ to the agents' behavior, which leads to the loss of surplus.*

[15]See Section 4.3 for the more detailed analysis of this effect.

[16]Total expected surplus in a matched pair is 130 points whereas total expected surplus in a non-matched pair is 70 points (the expected payoff of the agent).

## 4.3 Bargaining Power

In Section 4.1 we saw that principals earn significantly less than the theoretical predictions in both ID and MD, even conditional on successful matching, and that principals seem to adjust to the choices of the agents since they have much more to lose from a failure of persuasion or incentivization. Further, in Section 4.2 we saw that principals seem to be content with settling for a matching rate (successful persuasion or incentivization) of about 50% in order to not give up more of the expected surplus in the successful persuasion or incentivization periods. The (PP)SPE of the two games are not consistent with such behavior. First, agents should accept any offer above the one that guarantees them an expected a priori payoff. Second, through backward induction, principals should correctly anticipate agents' behavior and offer the least amount of informativeness or transfer such that agents follow the recommendation or accept the transfer. As such, in theory, both games admit a first-mover advantage for principals. However, in practice, the payoff asymmetry between agents and principals and most notably the "threat" of the non-match for principals may introduce some implicit "bargaining power" for the agents. Thus, in this section we analyze the Nash Bargaining Solution (NBS) of the two games, which—unlike the non-cooperative equilibrium concepts—takes into account the outside options of the players.

To proceed with this analysis, notice that, conditional on a pair matching ($X \geq Y$), the share of the total expected surplus (130 points) that each player receives is uniquely determined by $X$, the principal's choice (see footnote 13). Therefore, it is in the principal's best interest to choose $X$ as low as possible, while it is in the agent's best interest to try to force the principal to choose a high $X$. Even though it is the principal's choice that determines the share of expected surplus for the two players, the agent can effectively threaten the principal with the possibility of a non-match by increasing $Y$, implicitly forcing the principal's choice upwards. Thus, while theoretically the first-mover advantage gives the principal full bargaining power, in practice the asymmetry in each player's outside option may change that by introducing behavioral asymmetries between the two players and across the two games.

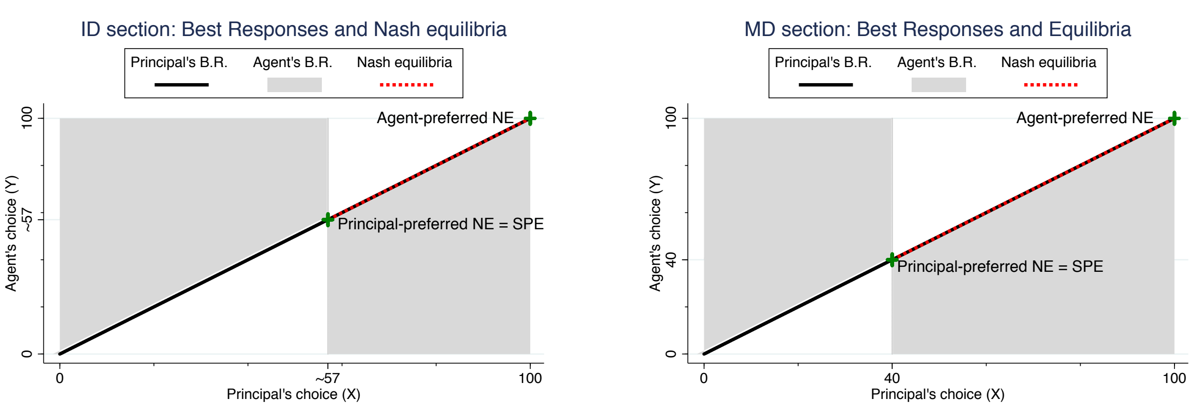



Figure 3: Best response correspondences and Nash Equilibria in the ID and MD games.

The questions then are What theoretical outcome should this bargaining process yield given the parameters of the ID and MD games? and Can the behavior of the participants be explained by some Nash Bargaining Solution? To find the answers we first look at the normal forms of the two games. The strategy sets of the two players are $X \in [0, 100]$ and $Y \in [0, 100]$. Figure 3 shows the best response correspondences and the sets of Nash equilibria in the ID and MD games (see Appendix A.3 for details). One can easily see that the best responses in the games are the same except for the point of the switch in the agent's best response correspondence. There is a continuum of NE that range from the agent-preferred to the principal-preferred, which is also the (PP)SPE of the extensive form game. Thus, the two games admit a similar "non-cooperative" structure.

Figure 4 shows the possible outcomes of the games in the expected payoff space (any choice of $X$ and $Y$ maps into some pair of expected payoffs). The black lines represent the possible expected payoffs in case there is a match ($X \geq Y$), and the "Non-match" points show the payoffs in case $X < Y$. Notice, however, that the disagreement outcomes are not necessarily the same as a non-match. We calculate them as the minimal expected payoffs that each player can guarantee regardless of the choices of the other. In the ID game the principal can guarantee herself 30 points by choosing $X = 100$, which the agent is forced to accept by the design of the game. In the MD game the principal can only guarantee herself 0 points by choosing $X = 100$. The agent can always get the minimum of 70 points by choosing $Y = 100$ in either game.[17] From the graphs it is clear that, given these disagreement outcomes, the Nash Bargaining Solution will predict the choice of one of the Nash equilibria described above, mapping 1-to-1 from the set of bargaining weights of the players (see Appendix A.3 for details).

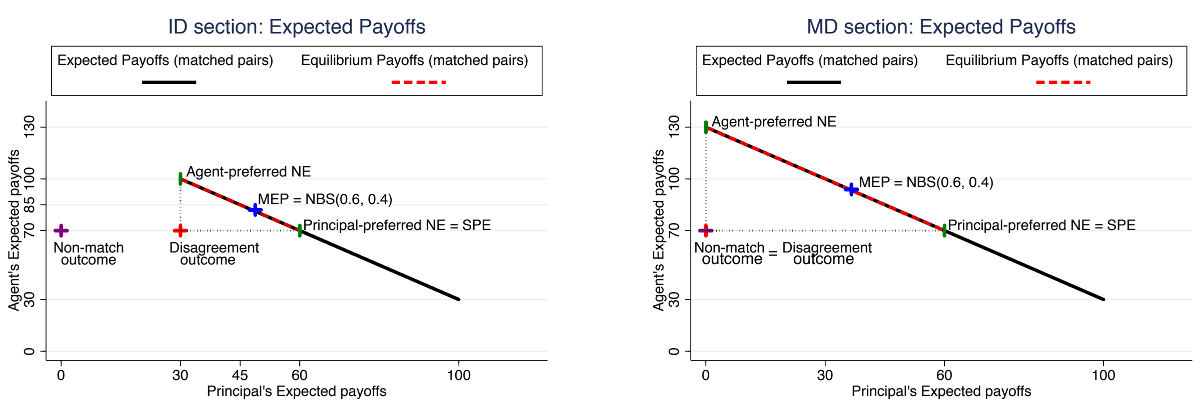



Figure 4: Possible expected payoffs in the ID and MD games, disagreement outcomes, average Matched Expected Payoffs (MEP), and Nash Bargaining Solution with bargaining weights 0.6 and 0.4.

To see if the behavior in both games can be explained by the Nash Bargaining Solution with some fixed bargaining power parameters we take only matched interactions and calculate the Matched Expected Payoffs (MEP) as we did above separately for each of the 13 independent groups of

[17]More specifically the agent can guarantee 70 points by choosing $Y \in [57.1, 100]$ in ID and $Y \in [40, 100]$ in MD.

participants.[18] Blue crosses on Figure 4 show the overall average MEP. By inverting the NBS mapping from bargaining weights to a Nash equilibrium and taking each average MEP as the corresponding Nash equilibrium in each game, we can retrieve the implied bargaining weights of each player in each game. We perform this calculation in Appendix A.4 and find the principals' average bargaining weights to be 0.593 and 0.610 in the ID and MD games respectively, a remarkably similar result (standard errors: 0.09 and 0.06). This suggests two independent results. First, as expected, we find that agents are able to capture a relatively large part of the additional surplus (40%) in contrast to the theoretical predictions (0%). As we conjectured above, we attribute this result to the payoff asymmetry and more specifically to the threat of a non-match outcome that principals face. Second, the implied distribution of bargaining power is almost identical in the two games. By accounting for the difference in the outside options (non-match outcomes), the NBS can account for the differences in expected payoffs in the two games.

**Result 3.** *A large part of the surplus created by successful persuasions or incentivizations is captured by the agents in sharp contrast to the theoretical predictions. In fact, analyzing our results using the Nash Bargaining Solution and accounting for each player's minimum guaranteed expected payoff (outside option), we find the implied bargaining power distributions to be almost identical in the two games (60% - principals, 40% - agents). As such, while bargaining power is a driver of a large wedge between theory and practice, it does not seem to create any wedges between the informational and monetary environments.*

# 5 Conclusion

This paper utilizes a lab experiment to explore the behavioral side of a recently proposed parallelism between the information design (ID) and the more traditional theory of mechanism design (MD) (see Bergemann and Morris, 2017; Taneva, 2017). We modify the framework of Kamenica and Gentzkow (2011) to allow for the implementation of both the ID and MD principal-agent problems as games with identical action spaces, equivalent best response correspondences, and the same predicted expected payoffs for each player in the Principal-Preferred Subgame-Perfect Equilibrium [(PP)SPE]. The latter equivalence serves as our main object of comparison and the starting point for our behavioral analysis.

In order to minimize various contextual influences in our experiment, we have designed the ID and MD games in such a way that both principals and agents choose a single number in the same range in very similar and relatively simple environments. In both games there are two states of the world determined by the color of a ball drawn from an urn with common knowledge of the ball composition (seven blue balls and three red balls). The agent receives a reward if she guesses the color of the ball correctly, while the principal receives a reward if the agent guesses the color red.

[18] We discard the non-matches because they happen due to noise and miscoordination, whereas NBS assumes that players can choose to match.

In MD the principal chooses the amount of money to transfer to the agent for guessing red. In ID the principal chooses the informativeness of a recommendation sent to an agent about the guess of the color of the ball. In both games, the agents choose the minimum amount of monetary transfer (MD) or recommendation informativeness (ID) that they require in order to accept the transfer (MD) or follow the recommendation (ID). The similarity of the two choice environments allows us to make direct comparisons of the behavioral differences between the informational (persuasion) versus the monetary (incentives) environments.

The predictions of Bayesian persuasion are strongly supported in our data in terms of participants' average choices in the ID game. While individual behavior can be erratic and history-dependent, on average, both principals' and agents' choices are not statistically different from the theory in all 10 periods of the game.

Overall, our principals were able to use both informative recommendations and monetary transfers to either persuade or incentivize their respective paired agents to take an otherwise (ex-ante) undesirable action. Thus, we can establish that, at some level, both information design and mechanism design are viable strategies that our principals successfully use to some extent to improve upon their ex-ante expected payoffs.[19]

In our experiment, the strategy of persuasion (Information Design) proved to be more profitable for principals over the more traditional strategy of monetary incentives (Mechanism Design), despite the equivalence predicted by the theory. This result seems robust since principals are more often able to convince their paired agents to follow their recommendations in ID than accept their transfer in MD but even conditional on a successful persuasion or incentivization, persuasion attempts were more profitable for the principals. These results are strongly influenced by agents' higher demand for monetary compensation than the equivalent recommendation informativeness. As such, we can say that agents were "easier" to persuade than incentivize.

In both games, the principals' average earnings were significantly less than the prediction of Kamenica and Gentzkow (2011)'s Principal-Preferred Subgame Perfect Equilibrium [PP]SPE. This is partly because on many occasions (about half) the principals failed to persuade or incentivize their paired agents to follow the recommendation or accept the transfer. As a result, the agents ignore the recommendation or reject the transfer. Motivated by the strong asymmetry in the baseline payoffs (if the principal fails to persuade/incentivize the agent, the principal is guaranteed zero points while the agent still expects 70 points on average) we conjecture that the principals' fear of a zero payoff combined with the agents' security of 70 points would induce some form of implicit bargaining between the players. We analyze the Nash bargaining solutions in the two games, which include the prediction of the [PP]SPE as a special case with full bargaining power attributed to the principal. Our analysis showcases two results. First, contrary to the predictions of the [PP]SPE, agents are able to capture a large part of the surplus generated by successful persuasions and incentivizations. Second, we find that the implied distributions of bargaining power between principals and agents are

[19] In the absence of either of those strategies, the principal is guaranteed a payoff of zero.

similar in the two environments. Thus, while bargaining power creates a large gap between theory and practice, it does not drive any wedges between the informational and monetary environments.

## References


Alonso, R. and Câmara, O. (2016a). Bayesian persuasion with heterogeneous priors. *Journal of Economic Theory*, **165**, 672–706.

— and Câmara, O. (2016b). Persuading voters. *American Economic Review*, **106** (11), 3590–3605.

— and Câmara, O. (2016c). Political disagreement and information in elections. *Games and Economic Behavior*, **100**, 390–412.

Babichenko, Y. and Barman, S. (2016). Computational aspects of private bayesian persuasion. *arXiv preprint arXiv:1603.01444*.

Bergemann, D. and Morris, S. (2016). Information design, bayesian persuasion, and bayes correlated equilibrium. *American Economic Review*, **106** (5), 586–91.

— and — (2017). Information design: A unified perspective.

Bizzotto, J., Rudiger, J. and Vigier, A. (2016). *Dynamic bayesian persuasion with public news*. Tech. rep., Mimeo.

Boleslavsky, R. and Kim, K. (2018). Bayesian persuasion and moral hazard.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica*, pp. 1431–1451.

DellaVigna, S. and Gentzkow, M. (2010). Persuasion: empirical evidence. *Annu. Rev. Econ.*, **2** (1), 643–669.

Dughmi, S., Kempe, D. and Qiang, R. (2016). Persuasion with limited communication. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, ACM, pp. 663–680.

— and Xu, H. (2016). Algorithmic bayesian persuasion. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, ACM, pp. 412–425.

— and — (2017). Algorithmic persuasion with no externalities. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, ACM, pp. 351–368.

Fréchette, G. R., Lizzeri, A. and Perego, J. (2022). Rules and commitment in communication: An experimental analysis. *Econometrica*, **90** (5), 2283–2318.

Gentzkow, M. and Kamenica, E. (2014). Costly persuasion. *American Economic Review*, **104** (5), 457–62.

— and — (2016). Competition in persuasion. *The Review of Economic Studies*, **84** (1), 300–322.

— and — (2017). Bayesian persuasion with multiple senders and rich signal spaces. *Games and Economic Behavior*, **104**, 411–429.

Gratton, G., Holden, R. and Kolotilin, A. (2017). When to drop a bombshell. *The Review of Economic Studies*, **85** (4), 2139–2172.

Hernández, P. and Neeman, Z. (2018). How bayesian persuasion can help reduce illegal parking and other socially undesirable behavior.

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

Kolotilin, A., Mylovanov, T., Zapechelnyuk, A. and Li, M. (2017). Persuasion of a privately informed receiver. *Econometrica*, **85** (6), 1949–1964.

Li, F. and Norman, P. (2018). On Bayesian persuasion with multiple senders. *Economics Letters*, **170**, 66–70.

Mathevet, L., Perego, J. and Taneva, I. (2020). On information design in games. *Journal of Political Economy*, **128** (4), 1370–1404.

Okuno-Fujiwara, M., Postlewaite, A. and Suzumura, K. (1990). Strategic information revelation. *The Review of Economic Studies*, **57** (1), 25–47.

Taneva, I. (2017). Information design.

Wang, Y. (2013). Bayesian persuasion with multiple receivers. *Available at SSRN 2625399*.

# Appendix (for online publication)

# A Theoretical Derivations

## A.1 Equilibrium in the Information Design Extension

For this game we will make use of the Kamenica and Gentzkow's Principal-Preferred Subgame Perfect Equilibrium and solve the game through backward induction. We provide the intuition behind the solution, omitting detailed proofs which are provided in Kamenica and Gentzkow (2011) and other sources.

**Stage 2.** The agent's problem is to choose an action $c(m) \in \{r, b\}$ for every possible message $m \in \{\rho, \beta\}$ to maximize her expected payoff given the prior probability distribution $p$ over the states and the principal's signal structure $(P_R, P_B)$:

$$c^*(m) = \text{argmax}_{c\in\{r,b\}} \; P(m)\Pi^A_{R,c} + (1 - P(m))\Pi^A_{B,c}$$

where $P(m)$ denotes the posterior probability of state $R$ given the message $m$, which is generated from the principal's signal structure $(P_R, P_B)$—chosen by the principal in stage 1—and is calculated according to the Bayes' rule as follows:[1]

$$\begin{aligned} P(m) = \Pr(R|m) &= \frac{\Pr(m|R)\Pr(R)}{\Pr(m)} = \frac{p\Pr(m|R)}{p\Pr(m|R) + (1-p)\Pr(m|B)} \\ 1 - P(m) = \Pr(B|m) &= \frac{\Pr(m|B)\Pr(B)}{\Pr(m)} = \frac{(1-p)\Pr(m|B)}{p\Pr(m|R) + (1-p)\Pr(m|B)} \end{aligned}$$

Essentially, the agent Bayes-updates her beliefs about the likelihood of each state and then takes the action which matches the state that is more likely to have been realized under each message:

$$c^*(m) = \begin{cases} r, & \text{if } P(m) \geq 1/2. \\ b, & \text{if } P(m) < 1/2. \end{cases}$$

Note that we follow Kamenica and Gentzkow (2011) in resolving the indifference case (where $P(m) = 1/2$) by having the agent choosing the principal-preferred action $r$ (which is where the "principal-preferred" part in the equilibrium name comes from).

**Stage 1.** The principal's problem is to choose a signal structure $(P_R, P_B)$ to maximize her expected payoff, which is the probability with which the agent will choose action $r$ times the payoff derived from that action, $\Pr(c^*(m) = r)\Pi$. Given the agent's optimal behavior derived in Stage 2, this problem reduces to maximizing $\Pr[P(m) \geq 1/2]$. The intuition behind the solution goes as follows. Since $p < \frac{1}{2}$, it is impossible to have $P(m) \geq \frac{1}{2}$ for both $m \in \{\rho, \beta\}$ so the best that the principal can do is to choose one message for which the induced posterior will be weakly greater than half and for the other message strictly less than half. Without loss of generality and to facilitate the parallelism of messages as action recommendations we assume that the principal will choose to induce the posteriors such that $P(\rho) \geq \frac{1}{2}$, $P(\beta) < \frac{1}{2}$. (i.e., such that the agent will want to choose $c^*(\rho) = r$, $c^*(\beta) = b$). The principal thus seeks to maximize $\Pr(m = \rho)$ (equivalently, maximize $\Pr(m = \rho|s = R)$ and minimize $\Pr(m = \beta|s = B)$) while being constrained by $P(\rho) \geq \frac{1}{2}$, $P(\beta) < \frac{1}{2}$. To do so, the principal chooses to always transmit the correct message (recommendation) in the state where both players' preferred actions coincide (when $s = R$) and mix the recommendations in the state where the players' preferred actions conflict (when $s = B$) such that the following is satisfied:

[1] $P_R = \Pr(m = \rho|s = R)$, $P_B = \Pr(m = \beta|s = B)$.

$$\begin{aligned}
P(\rho) &\geq 1/2 \\
\frac{\Pr(m=\rho|s=R)\Pr(R)}{\Pr(r)} &\geq 1/2 \\
\frac{\Pr(m=\rho|s=R)p}{\Pr(m=\rho|s=R)p+\Pr(m=\rho|s=B)(1-p)} &\geq 1/2 \\
\frac{p}{p+\Pr(m=\rho|s=B)(1-p)} &\geq 1/2 \\
\Pr(m=\rho|s=B) &\leq \frac{p}{1-p} \\
\Rightarrow P_B = \Pr(m=\beta|s=B) &\geq \frac{1-2p}{1-p}.
\end{aligned}$$

Since the principal seeks to minimize $P_B$, the solution to the principal's problem is given by:

$$\begin{aligned}
P_R^* &= \Pr(m=\rho|s=R) = 1 \\
P_B^* &= \Pr(m=\beta|s=B) = \frac{1-2p}{1-p}
\end{aligned}$$

**Principal-Preferred Subgame Perfect Equilibrium of Information Design Game.** The principal's optimal signal structure derived above induces the following posteriors:

$$P(m) = \begin{cases} \frac{1}{2}, & \text{if } m=\rho \\ 0, & \text{if } m=\beta. \end{cases}$$

Correspondingly, the agent's optimal action-choice rule dictates the following choice rule in equilibrium:

$$c^*(m) = \begin{cases} r, & \text{if } m=\rho \\ b, & \text{if } m=\beta. \end{cases}$$

Given the above, each player's expected payoffs are given by:

$$\begin{aligned}
E_s\Pi^P_{s,c^*(m)} &= p\Pr(m=\rho|s=R)\Pi + (1-p)\Pr(m=\rho|s=B)\Pi \\
&= 2p\Pi \\
E_s\Pi^A_{s,c^*(m)} &= p\Pr(m=\rho|s=R)\Pi + (1-p)\Pr(m=\beta|s=B)\Pi \\
&= (1-p)\Pi
\end{aligned}$$

## A.2 Equilibrium in the Mechanism Design Extension

For this game, we again use the Principal-Preferred Subgame Perfect Equilibrium (where "Principal-Preferred" is again used to indicate that we will resolve the indifference case in favor of the principal) as the solution concept for this game.

**Stage 2.** The agent's problem is to choose an action $c \in \{r, b\}$ to maximize her expected payoff given the prior probability distribution $p$ over the states and the principal's transfers $(t_r, t_b)$:

$$c^* = \text{argmax}_{c\in\{r,b\}} \; p\Pi^A_{R,c} + (1-p)\Pi^A_{B,c} + t_c,$$

where $t_c$ denotes the action-contingent transfer that the principal chooses in stage 1. For risk-neutral expected utility maximizing agent, the optimal action is given by

$$c^* = \begin{cases} r, & \text{if} \quad t_r - t_b \geq (1-2p)\Pi \\ b, & \text{if} \quad t_r - t_b < (1-2p)\Pi. \end{cases}$$

Note that once again here we resolve the indifference case when $t_r = (1-2p)\Pi + t_b$ by having the agent choosing action $r$. Also note that in this case, given the principal's action-contingent transfers, the agent's action is deterministic. This is in contrast to the information design case where the agent's action-choice rule is a function of the probabilistically generated message.

**Stage 1.** The principal's problem is to choose (non-negative) action-contingent transfers $(t_r, t_b)$ in order to maximize her expected payoff, which is given by the payoff in the baseline game minus the transfer conditional on the agent's action: $\Pi^P_{s,c} - t_c$. Trivially, it is optimal to set $t^*_b = 0$ since $\Pi^P_{s,b} = 0, \forall s = \{R, B\}$. The problem then reduces to minimizing $t_r$ so that the agent finds it optimal to choose $r$:

$$\begin{aligned} t^*_r = \text{argmax}_{t_r \in [0,100]} \; & \Pi - t_r \\ \text{s.t.} \;\; & t_r \geq (1-2p)\Pi. \end{aligned}$$

The solution to the principal's problem is given by

$$\begin{aligned} t^*_r &= (1-2p)\Pi \\ t^*_b &= 0 \end{aligned}$$

**Principal-Preferred Subgame Perfect Equilibrium of Mechanism Design Game.** The principal's optimal transfer choice induces the agent to choose $r$ in equilibrium.

$$c^* = r$$

Given this, the player's expected payoffs are given by:

$$\begin{aligned} E_s\Pi^P_{s,r} &= \Pi - t^*_r &= 2p\Pi \\ E_s\Pi^A_{s,r} &= p\Pi + t^*_r &= (1-p)\Pi \end{aligned}$$

## A.3 ID and MD in Normal Form: Nash Equilibria and Bargaining

We find Nash equilibria by explicitly defining the best response correspondences. Principals' best response correspondences:

$$\begin{aligned} BR^{Principal}_{ID}(Y) = Y, \;\; \forall\, Y \in [0,100] \\ BR^{Principal}_{MD}(Y) = Y, \;\; \forall\, Y \in [0,100] \end{aligned}$$

Agents' best response correspondences:

$$BR^{Agent}_{ID}(X) = \begin{cases} (X, 100], & \text{if} \quad X < \frac{400}{7} \\ [0,\ 100], & \text{if} \quad X = \frac{400}{7} \\ [0, X], & \text{if} \quad X > \frac{400}{7} \end{cases}$$

$$BR_{MD}^{Agent}(X) = \begin{cases} (X, 100], & \text{if } \; X < 40 \\ [0,\ 100], & \text{if } \; X = 40 \\ [0, X], & \text{if } \; X > 40 \end{cases}$$

The intersection of the above best response correspondences identifies a continuum of Nash equilibria for each game as shown in Figure 5.

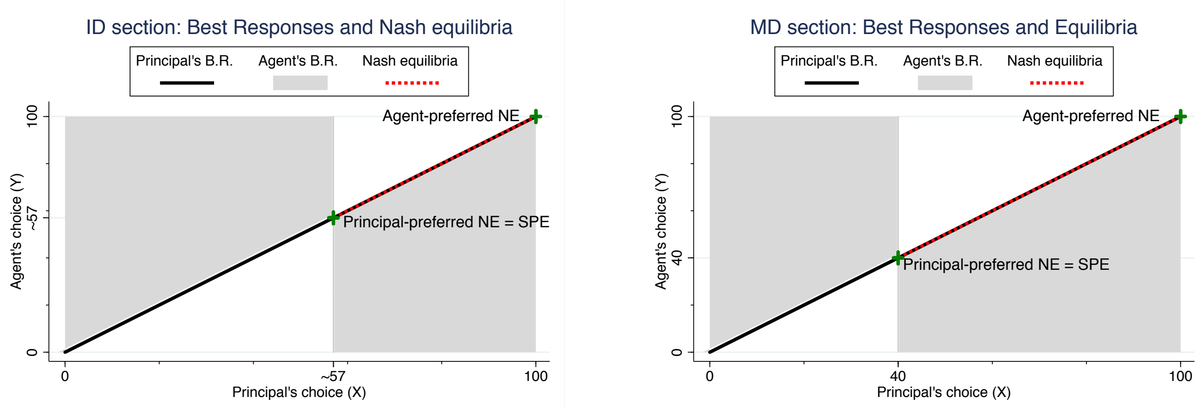



Figure 5: Best responses and Nash equilibria for the ID and MD games. "Agent-preferred NE" refers to the Nash equilibrium that is best for the agent in terms of expected payoffs. Correspondingly "Principal-preferred NE" refers to the Nash equilibrium that is best for the principal.

The set of Nash equilibria is very similar in the two games, except that the range is slightly larger in the MD game due to the fact that the (PP)SPE is at a lower point. The Principal-preferred Nash equilibria should not to be confused with the corresponding Principal-Preferred Subgame Perfect Equilibria, except that they coincide in both games. This merely reflects the fact that the first-mover advantage of the principal in the two-stage version of the games, assigns full bargaining power to the principal.

To highlight the equivalence of the NE of the two games, Figure 6 graphs them in the expected-payoff space. We note the following. Conditional on agreement ($X \geq Y$), no equilibrium outcome Pareto dominates any other (i.e., the surplus generated by persuasion or incentives does not depend on who receives it). Moving along the red striped line from left to right, the expected payoffs of the Nash equilibria increase for the principal and decrease for the agent in a linear fashion. Thus, conditional on agreement, both the ID and MD games resemble constant-sum games. The equilibrium outcomes predicted by the (PP)SPE corresponds to the best Nash equilibrium for the principal (Principal-preferred NE). We call "Disagreement outcome" the minimum guaranteed expected payoff that each player can guarantee in each game. This happens when the principal chooses $X = 100$ (guaranteed to persuade/incentivize) and when the agent chooses $Y \geq \frac{400}{7}$ in ID and $Y \geq 40$ in MD. We call "Non-match outcome" the expected payoffs for each player when the principal fails to persuade/incentivize ($X < Y$). While the disagreement outcomes are the same as the non-match outcomes for agents across the two games, it is not the case for principals. This is because by choosing $X = 100$, in ID the principal constructs a fully-informative signal structure which guarantees her 30 points in expectation (since the agent will choose $c^* = r$, 30% of the time) while in MD the principal transfers 100 points (all of her points) to the agent, thus essentially guaranteeing herself zero points.

Next, we cahracterize the Nash Bargaining Solution (NBS). For the ID game we solve the following constrained optimization problem:

$$\max_{P,A} \; (P-30)^{\alpha}(A-70)^{1-\alpha}$$
$$\text{s.t.} \;\; P + A = 130,$$

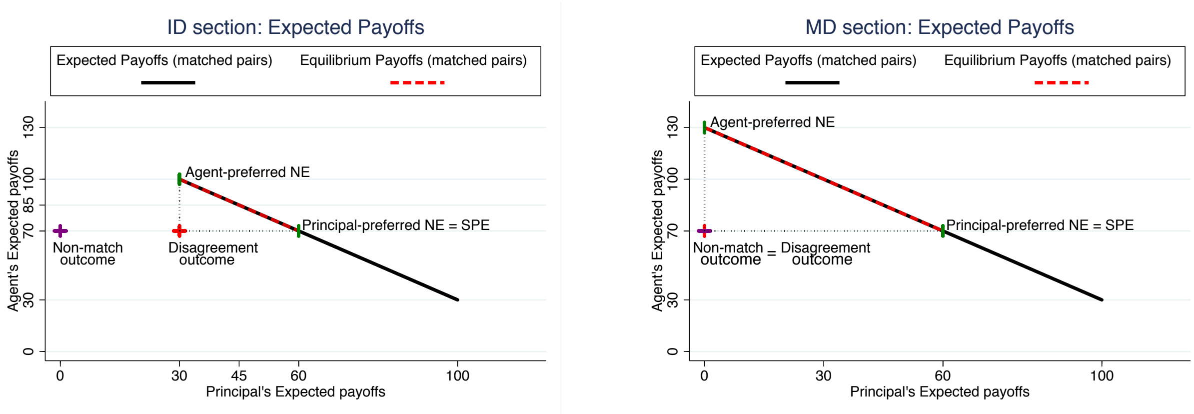



Figure 6: Expected payoffs in the ID and MD games. "Disagreement outcome" refers to the minimum guaranteed expected payoffs for each player. "Non-match outcome" refers to the expected payoffs when the principal fails to persuade/incentivize the agent ($X < Y$).

where $P$ and $A$ represent the principal's and agent's NBS agreement payoffs. 30 and 70 represent each player's disagreement outcomes while 130 is the total surplus. Finally $\alpha$ denotes the principals relative bargaining power. Since we, ex ante, take an agnostic view on the bargaining power we note that the above constrained optimization problem with equal bargaining power, $NBS(0.5, 0.5)$ admits a global maximum at the point $P = 45$, $A = 85$.

Similarly for the MD game, to find NBS we solve the following constrained optimization problem:

$$\max_{P,A} \ (P-0)^{\alpha}(A-70)^{1-\alpha}$$
$$\text{s.t. } \ P + A = 130,$$

where the only difference from the NBS in the ID game is that principal's minimum guaranteed expected payoff (disagreement outcome) decreases from 30 to 0. The $NBS(0.5, 0.5)$ admits a global maximum at the point $P = 30$, $A = 100$.

Thus, one can see that these two otherwise identical problems in terms of the (PP)SPE can predict a significant difference in expected payoffs (up to 50% less for the principal in MD than in ID) in terms of the Nash equilibria with uniform relative bargaining power for the two players. This happens because of the difference in the minimum guaranteed expected payoff that principals can guarantee themselves in ID and MD.

## A.4 Calculation of the Nash Bargaining Weights in ID and MD

In this section, we backtrack the relative Nash bargaining weights of each player from the observed matched expected payoffs in the ID game (left) and the MD game (right).

$$\max_{P^{ID},A^{ID}} \ (P^{ID}-30)^{\alpha^{ID}}(A^{ID}-70)^{1-\alpha^{ID}} \qquad \max_{P^{MD},A^{MD}} \ (P^{MD}-0)^{\alpha^{MD}}(A^{MD}-70)^{1-\alpha^{MD}}$$
$$\text{s.t. } \ P^{ID} + A^{ID} = 130 \qquad \text{s.t. } \ P^{MD} + A^{MD} = 130$$

The $P$ variables denote principal's expected payoffs and $A$ variables denote agent's expected payoffs in the corresponding game. From these constrained maximization problems we obtain the following relationships between principal's relative bargaining power $\alpha$ and the agent's matched average expected payoff in the

ID game (left) and MD game (right):

$$\alpha^{ID} = \frac{100 - A^{ID}}{30} \qquad\qquad \alpha^{MD} = \frac{130 - A^{MD}}{60}$$

Plugging the agents' average Matched Expected Payoffs (MEP) obtained from the data (see Section 4.3) we obtain the average principals' relative bargaining power in ID (left) and MD (right):

$$\alpha^{ID} = 0.593 \approx 0.6 \qquad\qquad \alpha^{MD} = 0.610 \approx 0.6$$

Interestingly, we observe that conditional on pairs matching, principals and agents exhibit similar relative bargaining powers across the two games. Taking this result at face value, it may appear that the difference in the absolute values of the Matched Expected Payoffs that we observe can be fully attributed to the difference in the minimum guaranteed outcome that principals can obtain in the two games.

# B Additional Analyses

| Section | Measure | Sample | Principals | | Agents | |
|---|---|---|---|---|---|---|
| | | | Data | Theory | Data | Theory |
| ID | Expected Payoffs | Matched pairs | 47.8<br>(2.73) | 60 | 82.2<br>(2.73) | 70 |
| MD | Expected Payoffs | Matched pairs | 36.6<br>(3.41) | 60 | 93.4<br>(3.41) | 70 |
| ID | Expected Payoffs | All pairs | 25.3<br>(1.7) | 60 | 76.1<br>(0.82) | 70 |
| MD | Expected Payoffs | All pairs | 16.4<br>(1.26) | 60 | 79.8<br>(0.91) | 70 |
| ID | Actual Payoffs | Matched pairs | 47.5<br>(3.67) | 60 | 81.5<br>(2.97) | 70 |
| MD | Actual Payoffs | Matched pairs | 38.1<br>(2.35) | 60 | 93.3<br>(4.01) | 70 |
| ID | Actual Payoffs | All pairs | 25.6<br>(2.36) | 60 | 76.9<br>(1.8) | 70 |
| MD | Actual Payoffs | All pairs | 16.4<br>(1.26) | 60 | 80.9<br>(2.23) | 70 |
| $N$ of observations | | | 13 | | 13 | |

Table 4: Summary of payoffs of Principals and Agents in the ID and MD sections split by full sample versus matched pairs only and expected versus realzed payoffs. Numbers in brackets indicate standard errors.

# C Instructions and Screenshots

All instructions below were translated from Italian since the experiment was run in Trento, Italy. For the Italian version, please contact the authors.

## C.1 Instructions for the ID Section

Section __ .

**General Information**

- Please read this instruction manual carefully as your understanding will play an important role in how much earnings you will make in this section. At the end you will be asked some comprehension questions to ensure your understanding. In addition, you will have the chance to play 2 practice rounds before the section starts.
- This section of the experiment you will play a simple game with one other person in this room. You will play this game **10 times (rounds) in a row.** Every round, you will be re-paired with a **different** person. One of you will act as a "**Player A**" and the other as a "**Player B**". Your role of either "Player A" or "Player B" will be determined randomly at the beginning of the section and you will keep the same role throughout this section.
- At the **end of each round**, you may earn some points. The amount will depend on your choices, the choices your paired participant and luck.
- At the **end of the section, one of the 10 rounds will be selected at random. Only the points you made in that round will count as your total points earned for this Section.**

**Overview of each round**

A ball will be randomly drawn from an urn containing **3 RED balls** and **7 BLUE balls.**

Player A's role will be to **recommend a guess** to Player B ("Guess red" or "Guess blue") depending on the color of the ball drawn.

Player B's role will be to **make a guess** (Red or Blue) by **following** or **ignoring** Player A's recommendation. Player A will earn 100 points if the guess is **Red**. Player B will earn 100 points if the guess is **correct**.

Each player's points are summarized below:

**Player A's points:**

| | If ball is **RED** | If ball is **BLUE** |
|---|---|---|
| If **Player B** guesses Red | 100 | 100 |
| If **Player B** guesses Blue | 0 | 0 |

**Player B's points:**

| | If ball is **RED** | If ball is **BLUE** |
|---|---|---|
| If **Player B** guesses Red | 100 | 0 |
| If **Player B** guesses Blue | 0 | 100 |

**Player A's choice in each round**

Player A will always give the correct recommendation to Player B if the ball is **RED** (i.e. "Guess red").

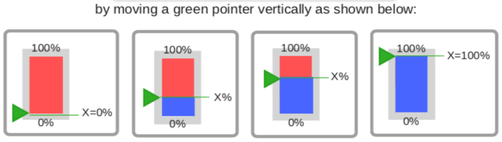
**Player A will choose the % chance [X%] of giving the correct recommendation to Player B if the ball is BLUE (i.e. "Guess blue")**

Thus, if the ball is **BLUE**, the recommendation can be either "Guess blue" or "Guess red".

Player A's recommendations will **automatically be generated** as follows:

| | Ball color: Red | Ball color: BLUE |
|---|---|---|
| Chance of recommendation "Guess red" | 100 % | 100 - X % |
| Chance of recommendation "Guess blue" | 0 % | X % |

**Player A will be able to choose X (0 to 100)**
by moving a green pointer vertically as shown below:




The leftmost example shows the case where a Player A will **never give the correct recommendation** if the ball is **BLUE** (always recommend "Guess red" if the ball is **BLUE**.)

As the examples progress from left to right, the the pointer is placed higher. Player A is choosing a larger X. i.e. A **higher % chance of giving the correct recommendation** if the ball is **BLUE** (lower % that the recommendation will be "Guess red" if the ball is **BLUE**.)

The rightmost example shows the case where Player A will **always give the correct recommendation** if the ball is **BLUE**. (never recommend "Guess red" if the ball is **BLUE**)

**Player B's choice in each round**

Player B will **always** receive the correct recommendation if the ball is **RED** ("Guess red"). Thus Player B will decide to **follow** or **ignore** Player A's recommendations based on how often Player A gives the correct recommendation when the ball is **BLUE** ("Guess blue").

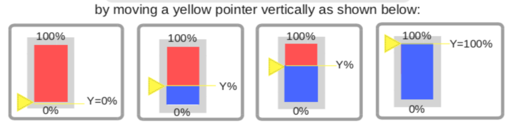
**Player B will choose the minimum % chance [Y%] of receiving the correct recommendation when the ball is BLUE that Player B is willing to accept in order to follow Player A's recommendations.**

Player B's **guess will automatically be determined** depending on
1. Whether Player B **follows** or **ignores** Player A's recommendations.
2. The recommendation generated.

The following table shows Player B's guess in each case:

| | | **Choices** of Player A & Player B: X ≥ Y | **Choices** of Player A & Player B: X < Y |
|---|---|---|---|
| Recommendation | "Guess red" | Red | Blue |
| | "Guess blue" | Blue | Blue |
| | | Player B automatically **follows** Player A's recommendations | Player B automatically **ignores** Player A's recommendations |

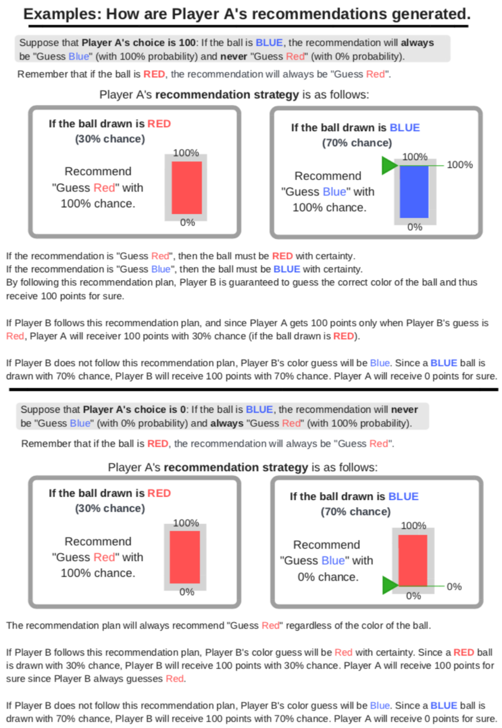
**Player B will be able to choose Y (0 to 100)**
by moving a yellow pointer vertically as shown below:

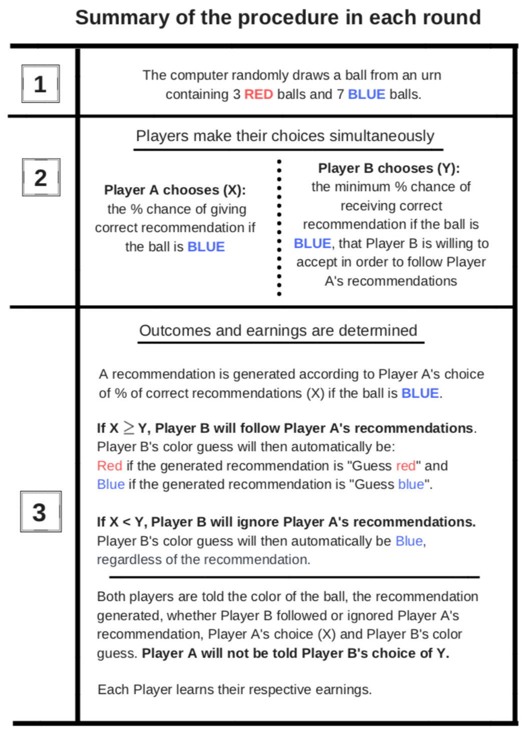



The leftmost example shows the case where a Player B is willing to follow Player A's recommendations for **any % chance of correct recommendations** when the ball is **BLUE**. (Since X cannot be less than 0).

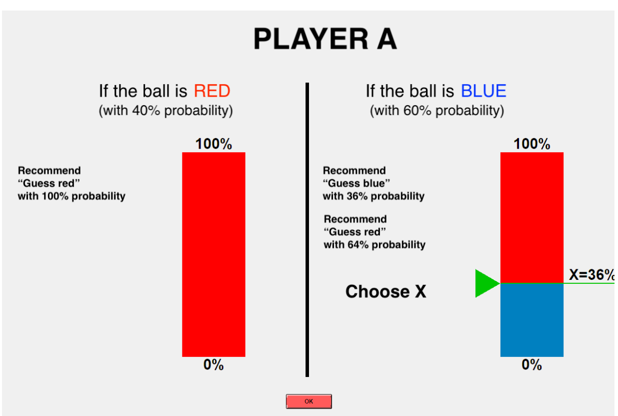
As the examples progress from left to right, the pointer is placed higher. Player B is choosing a larger Y. i.e. Only willing to follow Player A's recommendations with **higher % of correct recommendations** when the ball is **BLUE**. (Since X needs to be larger in order to satisfy Y)

The rightmost example shows the case where a Player B is only willing follow Player A's recommendations if **recommendations are always correct** when the ball is **BLUE**. (since X needs to equal 100 in order to satisfy Y)




## C.2 Screenshots for the ID section



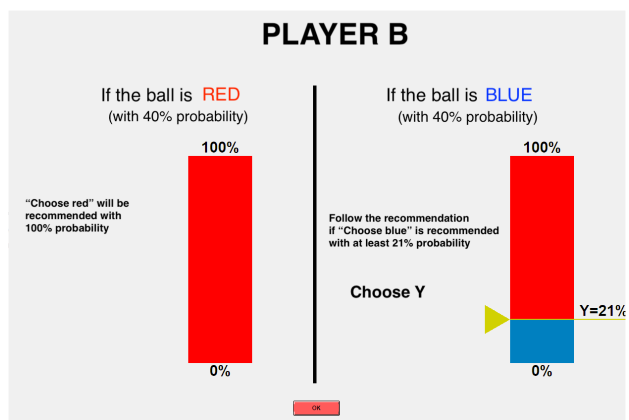




Figure 7: Choice screens of Player As (principals) and Player Bs (agents) in each period in the ID section.

## C.3 Instructions for the MD Section

### Section __ .

**General Information**

- Please read this instruction manual carefully as your understanding will play an important role in how much earnings you will make in this section. At the end you will be asked some comprehension questions to ensure your understanding. In addition, you will have the chance to play 2 practice rounds before the section starts.
- This section of the experiment you will play a simple game with one other person in this room. You will play this game **10 times (rounds) in a row.** Every round, you will be re-paired with a **different** person. One of you will act as a "**Player A**" and the other as a "**Player B**". Your role of either "Player A" or "Player B" will be determined randomly at the beginning of the section and you will keep the same role throughout this section.
- At the **end of each round**, you may earn some points. The amount will depend on your choices, the choices your paired participant and luck.
- At the **end of the section, one of the 10 rounds will be selected at random. Only the points you made in that round will count as your total points earned for this Section.**

**Overview of each round**

A ball will be randomly drawn from an urn containing **3 RED balls** and **7 BLUE balls.**

**Player A's role** is to **propose to transfer points** to Player B for guessing Red. Player A will transfer these points to Player B if Player B guesses Red.

**Player B's role** is to **make a guess** (Red or Blue). Player A will earn 100 points if the guess is **Red**. Player B will earn 100 points if the guess is **correct**.

The resulting points are summarized below:

**Player A's points:**

| | If ball is **RED** | If ball is **BLUE** |
|---|---|---|
| If **Player B** guesses Red | 100 - Transfer | 100 - Transfer |
| If **Player B** guesses Blue | 0 | 0 |

**Player B's points:**

| | If ball is **RED** | If ball is **BLUE** |
|---|---|---|
| If **Player B** guesses Red | 100 + Transfer | Transfer |
| If **Player B** guesses Blue | 0 | 100 |

**Player A's choice in each round**

Player A will **never** transfer points to Player B if Player B guesses Blue.

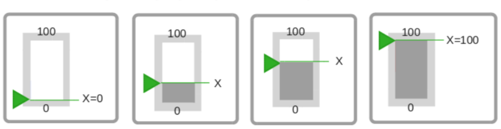
**Player A will choose the number of points [X] to transfer to Player B if Player B guesses Red.**

Thus, if Player B's guess is Red, Player A will transfer X points to Player B.

Player A's transfer will **automatically be executed** as follows:

| | Ball color: RED | Ball color: BLUE |
|---|---|---|
| Player A's **transfer** to Player B: | X | 0 |
| Player A's **remaining points**: | 100 - X | 0 |

**Player A will be able to choose X (0 to 100)**
by moving a green pointer vertically as shown below:




The leftmost example shows the case where a Player A chooses **not transfer any points** to Player B (keep all points) if Player B guesses Red.

As the examples progress from left to right, the the pointer is placed higher. Player A is choosing a larger X. ie. To **transfer more points** to Player B (keep less points) if Player B guesses Red.

The rightmost example shows the case where Player A chooses to **transfer all 100 points** to Player B (keep no points) if Player B guesses Red.

**Player B's choice in each round**

Player B will never receive a transfer from Player A, if Player B guesses Blue. Thus Player B will decide to **accept** or **reject** Player A's transfer based on the number of points that Player A transfers if Player B guesses Red.

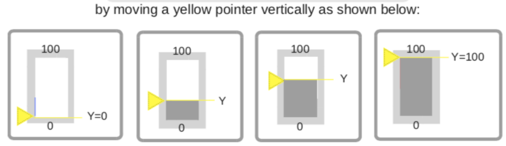
**Player B will choose the minimum transfer of points [Y] from Player A that Player B is willing to accept in order to guess Red.**

Player B's **guess will automatically be determined** depending on whether Player B **accepts** or **rejects** Player A's proposed transfer of points.

The following table shows Player B's guess in each case:

| **Choices** of Player A & Player B | $X \geq Y$ | $X < Y$ |
|---|---|---|
| Player B's **guess**: | Red | Blue |
| | Player B automatically **accepts** Player A's transfer. | Player B automatically **rejects** Player A's transfer. |

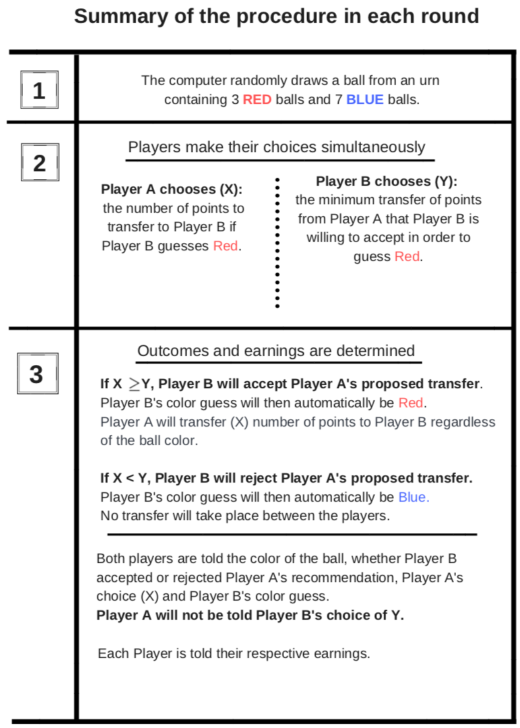
**Player B will be able to choose Y (0 to 100)**
by moving a yellow pointer vertically as shown below:




The leftmost example shows the case where a Player B is **willing to accept any transfer** of points in order to guess Red. (Since X cannot be less than 0)

As the examples progress from left to right, the pointer is placed higher. Player B is choosing a larger Y. i.e. Only **willing to accept transfers with more points** from Player A in order to guess Red. (Since X needs to be larger in order to satisfy Y)

The rightmost example shows the case where a Player B is **only willing to accept a transfer of all 100 points** from Player A in order to guess Red. (Since X needs to equal 100 to satisfy Y)




## C.4 Screenshots for the MD section

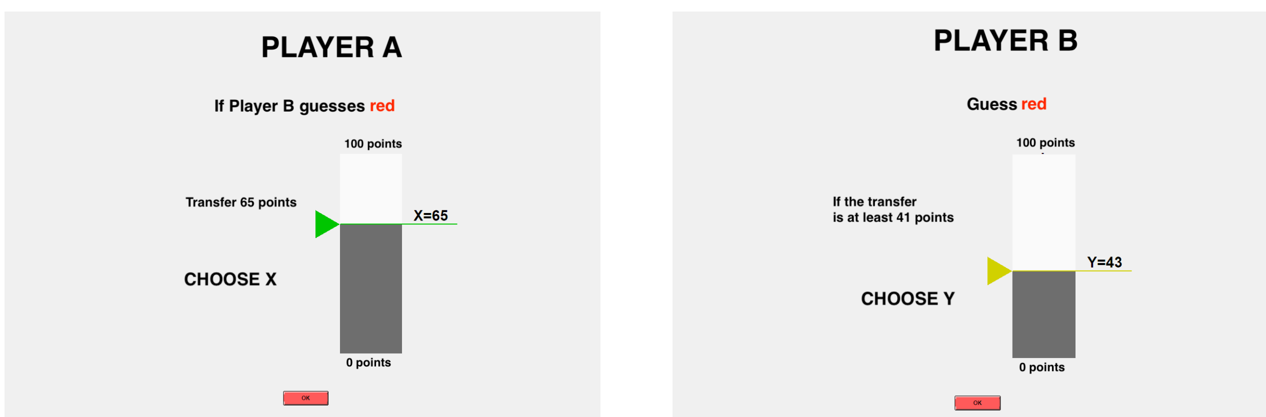



Figure 8: Choice screens of Player As (principals) and Player Bs (agents) in each period in the MD section.