Incentives or Persuasion? An Experimental Investigation

Maastricht University

Andreas Aristidou,
Giorgio Coricelli,
Alexander Vostroknutov

Incentives or Persuasion? An Experimental Investigation

RM/19/012

GSBE

Maastricht University School of Business and Economics
Graduate School of Business and Economics

P.O Box 616
NL- 6200 MD Maastricht
The Netherlands

Incentives or Persuasion? An Experimental Investigation*

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

March 25, 2019

Abstract

There are two theoretically parallel ways in which principals can manipulate agents' choices: with monetary incentives (mechanism design) or Bayesian persuasion (information design). We are interested in whether incentives or persuasion is a better strategy for principals. We conduct an experiment that investigates the behavioral side of the theoretical parallelism between these approaches. We find that principals are more successful when persuading than when incentivizing. Agents appear to be more demanding in mechanism design than in information design. Our analysis also identifies many features that make mechanism and information design behaviorally distinct in practice.

JEL classifications: *C91, C92, D47, D91.*
Keywords: *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, Philippos Louis, Tomaz Sadzik, Joel Sobel, Omer Tamuz, Nikolaos Tsakas, Joel Watson. This paper benefited greatly from discussions at the 93rd WEAI annual conference and the brown-bag seminars at the Department of Economics at USC. All errors are our own.

[†]University of Southern California (aaristid@usc.edu) - Corresponding author.
[‡]University of Southern California (coricell@usc.edu).
[§]Maastricht University (a.vostroknutov@maastrichtuniversity.nl).

1 Introduction

There are two strategies that a principal can employ to induce an agent to take a particular action. One possibility is to use monetary incentives that impact the agent's payoffs, the other is through Bayesian persuasion, a strategy that affects the agent's beliefs about the likelihood of outcomes following her choice. 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 product being good and acts to maximize his expected utility. What do online retailers usually 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 informational persuasion, for example, personalized recommendations, expert reviews, or product placement, in order to change the consumers' beliefs about the quality of the product. While in the past monetary incentives were the main form of motivating economic agents, today the manipulation of information becomes the force to reckon with. The emergence of massive information gatekeepers who own and use information to guide consumer behavior is shifting the balance of power from incentives to informational persuasion.

The goal of our study is to conduct a comparative analysis, both theoretical and behavioral, of the two strategies that principals can use to influence agents' choices. We draw our inspiration from the recent literature that attempts to study Bayesian persuasion in a general game-theoretic framework (Kamenica and Gentzkow, 2011), making parallels with mechanism design and naming it "Information Design."[1] 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. They consider three ingredients of a game: (1) the basic game structure (the sequence of moves, action spaces); (2) the payoff structure and (3) the information that agents possess regarding payoff-relevant states and other players' types. The game structure (1) is assumed given. An information/mechanism designer takes (2)/(3) as given and optimizes the design of (3)/(2). While, as is conventional in mechanism design, the designer is assumed to have the ability to commit to a transfer mapping to the players, 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.[2]

[1]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, 2017; Wang, 2013). The early seminal papers include Crawford and Sobel (1982) and Okuno-Fujiwara *et al.* (1990).

[2]For another interesting approach see Mathevet *et al.* (2017).

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, this raises several natural questions. How does this theoretical parallelism play out in practice? Can information designers use persuasion with the same effect as mechanism designers use incentives? How do agents 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 investigate whether they are more successful in using incentives or persuasion to influence agent's choice and to increase their payoffs.

In the theoretical part of the paper we use the standard two-state Bayesian persuasion model (Kamenica and Gentzkow, 2011). We start with a baseline setup in which a principal is not able to act as information or mechanism designer and trivially show that, in this case, she is guaranteed zero 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 and 2) the principal can act as a mechanism designer in an attempt to incentivize the agent to take the principal-preferred action. Using the Kamenica and Gentzkow's "Principal-Preferred" Subgame Perfect Equilibrium we show that the two games are equivalent in terms of best response correspondences and that the expected payoffs of both the principal and the agent are identical in equilibrium.[3]

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) information design in its standard form is too computationally intensive for lab participants and 2) information design and mechanism design are very different tasks, thus in order to detect behavioral differences that arise exclusively from their inherent features, 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. Apart from allowing us to tackle the research questions stated above, our experiment is the first attempt to experimentally investigate the theoretical parallelism between information and mechanism design that also overcomes many of the challenges in bringing the theoretical setup of Bayesian persuasion to the lab (Fréchette *et al.*, 2018; Nguyen, 2016; Au and Li, 2018).

We find that principals are able to use persuasion (information design) to a much better extent than incentives (mechanism design) in order to manipulate agents to choose their preferred action. Principals are able to persuade agents more often than they are able to incentivize them, and successful persuasions are, on average, more profitable than successful incentivizations. This result

[3] 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*.

seems to be driven by the fact that agents are more demanding when they are being incentivized than when they are being persuaded, which hinges critically on agents' perception of the relative value of information and money. However, principals' payoffs in both games still fall short of the theoretical predictions of the Principal-Preferred Subgame Perfect Equilibrium, which according to our analysis ignores a very important aspect of the principal-agent interaction: the distribution of bargaining power. While the two-stage nature of the non-cooperative game attributes 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, a result that holds in both information and mechanism design tasks.

Participants' behavior in the information design task is close to the equilibrium in terms of average choices. Nevertheless, we find that heterogeneous and erratic individual decisions cause large inefficiencies and have detrimental effect on the principals' payoffs. Exploring the inherent differences between information and mechanism design, we find that the smoother payoffs that mechanism design admits (transfers between players mitigate extreme payoffs) make players more stable in their choices. Conversely, the extreme nature of payoffs in information design causes strong reactions that lead to inefficiencies.

We also find that the nature of the "contract" between agents and principals is perceived differently in information and mechanism design. When agents are successfully persuaded but end up with a bad outcome they exhibit an extreme reaction, reminiscent of betrayal aversion. In mechanism design such reaction does not arise since agents are directly compensated for exactly that contingency. Thus, belief manipulation can have behavioral side effects that are avoided with incentivization.

Our data suggest that participants have much more difficulties dealing with probabilistic reasoning in information design than with monetary transfers in mechanism design. Specifically, agents' choices are far from what they themselves believe is optimal to do in information design, but are very close to their beliefs in mechanism design.

Overall, our experiment shows that behavioral differences in participants' perception of informational and monetary incentives make the theoretical parallelism between mechanism and information design harder to motivate in practice. Thus, real-world designers should consider which strategy is most effective in each specific situation. While in our setting principals are better off persuading than incentivizing agents, it is likely that other environments will produce different results.

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. The framework is a version of the original two-state Bayesian persuasion setup from Kamenica and Gentzkow (2011). We begin with an adaptation of the model where the principal (sender) is stripped of his ability 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 (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	Π , Π	Π , 0
	b	0 , 0	0 , Π

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

$$E_s \Pi^P_{s,c^*} = 0 \qquad (\text{Principal}^{\text{Baseline}})$$

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

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 (ρ and β are principal's recommendations about the color of the ball, red or blue).[4] 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). Thus, 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:[5]

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

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

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 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

[4] A message ρ or β is always sent. When the ball is red, the message β is sent with probability $1 - P_R$ and when the ball is blue the message ρ is sent with probability $1 - P_B$.

[5] See Appendix A.1 for derivations.

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.[6] 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:[7]

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

Note the following: 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:

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

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.

3 Experiment

The experiment consisted of two treatments with three sections in each: Section ID (information design), Section MD (mechanism design), and Section 3. The two treatments differed in the order of sections ID and MD, while Section 3 was always implemented the last. 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

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

[7] See Appendix A.2 for derivations.

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.[8] Section 3 consisted of several tasks designed to help us to measure the behavioral traits potentially responsible for behavioral differences in the ID and MD games. 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 three choices in total).

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 by eliciting agent's choices with the strategy method. The principal constructed a signal structure (ID) or transfers (MD) and, at the same point in 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.* (2018), 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. We denote this choice by $X \in [0\%, 100\%]$. 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. When $X = 0\%$ the recommendation is always

[8] Feedback for principals and agents was not the same and was structured to reflect the information that each player would receive in an extensive form game. See Section 3.3 for details.

[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.* (2018) allow subjects to manipulate the equivalent of P_R and find that the vast majority of choices are at equilibrium ($P_R = 1$). We, thus, believe that our simplification does not significantly impact the behavior.

"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" ($P_R = 1$).

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 recommendation means that the agent guesses the color that the recommendation suggests. Ignoring the recommendation means that the agent guesses blue, 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. As a result, the agent's expected payoff from following a recommendation from $(1, X)$ is greater than that from following the recommendation from any $(1, X')$ with $X' > X$. By choosing Y in this manner the agent agreed ex ante to follow the recommendation coming from a signal structure that is at least as informative as $(1, Y)$ and to ignore the 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 follow only very informative signals, while agents who choose low Y are easy to persuade and follow recommendations for 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 information or transfer that would induce the agent to follow the recommendation or accept the transfer.

3.4 Summary of the Experimental Procedure

In each period (of 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 G 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 Section 3

Section 3 was always implemented the last (after sections ID and MD), and its purpose was to help us to uncover the behavioral traits responsible for the observed differences in behavior between ID and MD. It consisted of 4 choices in the order described below. All tasks were incentivized and one was chosen randomly for payment.

Dictator ID Participants were randomly matched in pairs to play one round of ID. Each player could fully control the outcome, which we call a Dictator choice. That is, principals chose X under the condition that $Y = 0$, and agents chose Y under the condition that $X = Y$ for any chosen Y. Thus, all pairs were forced to match with the most favorable conditions for the players who were choosing. This task served to control for differences in fairness considerations in the two games and was incentivized in the same way as all the choices in the ID games.

Dictator MD Same as above only with MD.

Cutoff ID Participants were incentivized to give their best estimate of the Bayesian rational cutoff in the ID game (the ID two-stage equilibrium outcome, 57.1 points). Specifically, they were paid proportionally to how close their answer was to the actual cutoff. This served to control for participants' beliefs about the rational play in the ID game.

Cutoff MD Same as above only with MD (the MD two-stage equilibrium outcome with risk neutral agent, 40 points).[11]

3.6 Design Implementation

The experiment was conducted in June 2018 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.

[11] To control for the possible effects of risk preferences we also elicited the MD cutoff unconstrained by risk neutrality. Since we did not find statistically significant differences between the two measures, in our analysis we use only the risk neutral one since our theory is based on risk neutral players.

Participants were informed that they will take part in an experiment with three 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?

We start our analysis by looking at the average choices X and Y and comparing them to the theoretical predictions. In order to be able to use the non-parametric tests we consider averages over choices in each of the 13 independent groups of participants described in Section 3.6. 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 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 tests, $p < 0.002$). This means that agents are, on average, more resistant to incentives than to the equivalent persuasion. In order to comply with the principal they ask for more points in MD than for the equivalent signal informativeness in ID. This can make it harder for principals to manipulate agents' choices through monetary incentives, thus potentially making persuasion a more successful strategy.

Section	Measure	Sample	Principals		Agents	
			Data	Theory	Data	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 independent 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. "Data" 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. We find that in ID principals are successful at persuasion 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, principals are more often successful when persuading rather than when incentivizing agents. However, we still only observe pairs matching about half of the time.[12]

Next, we look at the payoffs that principals and agents receive given a successful match. 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' 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.

Section	Measure	Sample	Principals		Agents	
			Data	Theory	Data	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 independent 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, principals still earn significantly less than the equilibrium prediction (signed-rank tests, ID: $p = .0037$, MD: $p = 0.0015$). These 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 yields more payoff for the principals than successful incentivization. This can be partially attributed to agents' demanding higher monetary transfers in MD than the equivalent informativeness of the recommendation in ID.*

[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 $(100 - X, 30 + X)$.

4.2 Evolution of Choices in Time

In order to understand what drives the aggregate results in the previous section we examine the per-period evolution of average choices in 13 independent groups shown in Figure 1. Two features are immediately noticeable. First, the average play is around the equilibrium in all 10 periods in ID, while generally higher than equilibrium in MD. Second, while choices in ID remain relatively stable for both principals and agents, in MD we observe principals starting at the equilibrium level in period 1, then gradually increasing their choices until they reach agents' average choices, which remain relatively stable throughout the game.

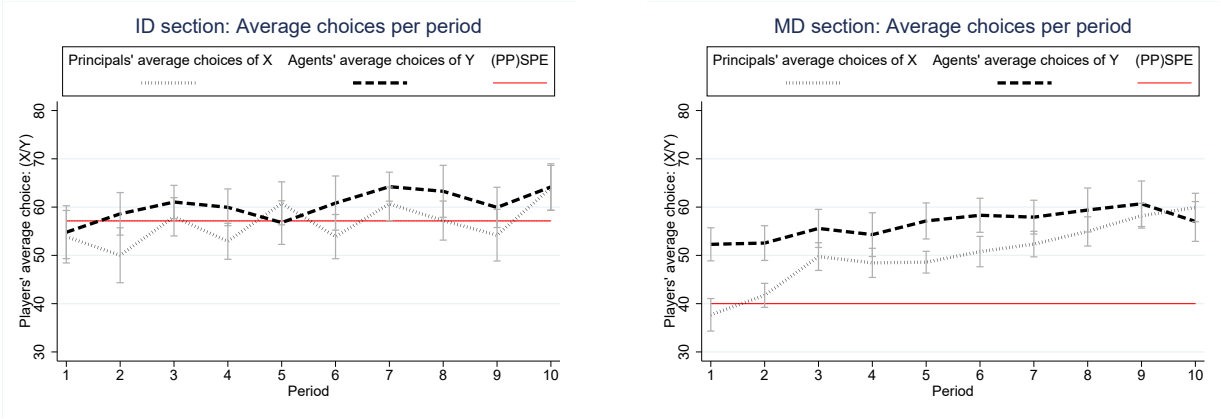


Figure 1: Average choices of principals (dotted lines) and agents (dashed lines) in each period of the ID and MD games. The solid red lines indicate the predictions of the two-stage (PP)SPE. Error bars are ±1SE corresponding to 13 observations.

This behavior may indicate the existence of an inherent stable level of agents' choices (different in each game) towards which principals converge. In ID, by chance or not, 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, whereas principals are guaranteed 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. 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 reaching the levels of ID matches around period 8. Notice that the overall percentage of matches is around 50% even in the late periods, which is extremely low.

Result 2. *Both games exhibit large inefficiencies in terms of the social surplus lost due to unsuccessful matches: about half of the pairs fail to match. While overall the situation is worse in MD, the difference decreases over time and in the last two periods the number of unsuccessful attempts is the same in the two games.*

Next, we look at the evolution of choices for pairs that match as shown in Figure 3. We observe large gaps between the average choices of principals and agents in both games that appear to be stable over time (about 25% in ID and 20 points in MD). Thus, principals are giving away a

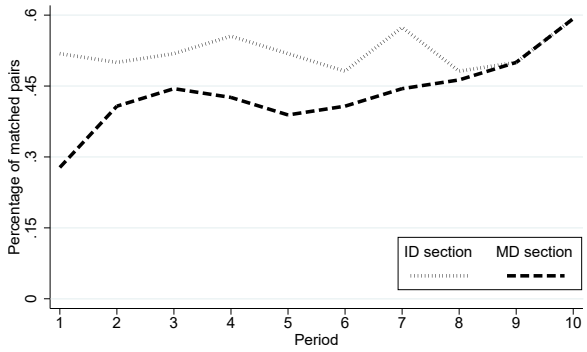


Figure 2: Percentage of matched pairs ($X \geq Y$) per period in each game.

much larger share of the surplus than what their paired agents are willing to accept, surplus they could have captured by lowering X.[14] We hypothesize that the main reason behind the size and the stability of the gaps is the heterogeneity in agents' choices coupled with the principals' sensitivity to unsuccessful persuasions/incentivizations (in which case they end up with zero points). In support of this idea remember that around half of the pairs do not match, which suggests that there is a tangible threat for principals who choose a low X. Moreover, the histograms of the agents' choices in Figure 11 in Appendix D show significant levels of very high choices by agents (more in ID than in MD). We explore these issues in more detail in Section 4.4 below.

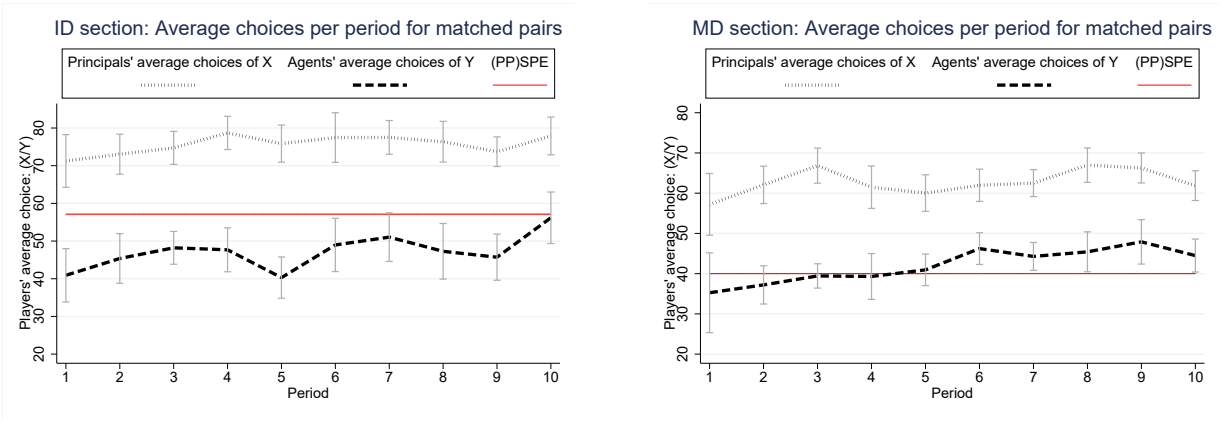


Figure 3: Average choices of matched pairs for each period of the ID and the MD games. The solid red lines indicate the predictions of the two-stage (PP)SPE. Error bars are ±1SE corresponding to 13 observations.

Result 3. *In an attempt to best respond to the heterogeneity in agents' behavior, principals increase their choices and give up a larger portion of the surplus than necessary. This is similar in both games.*

4.3 Nash Bargaining Solution

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

[14] The average gap of 25% in ID translates to about 18 points.

to the choices of the agents since they have much more to lose from a failure of persuasion or incentivization. Such behavior is not consistent with the (PP)SPE and especially with the fact that principals have the first mover advantage (agents' choices are still the reactions to their move even when chosen with the strategy method). However, our observations do suggest that what is important behaviorally is the relative "bargaining power" of the principals and 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.

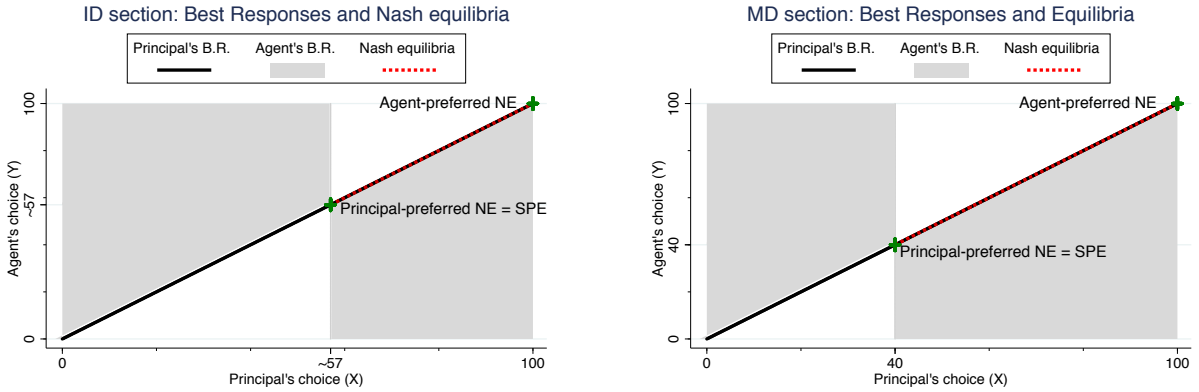


Figure 4: Best response correspondences and Nash Equilibria in the ID and MD games.

The questions now 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 4 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 games are the same in terms of their "non-cooperative" structure.

Figure 5 shows the possible outcomes of the games in the expected payoffs 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.[15] From the graphs it is clear that, given these disagreement outcomes, Nash bargaining solution predicts the choice of one of the Nash equilibria described above depending on the bargaining power of the players (see Appendix A.3 for details).

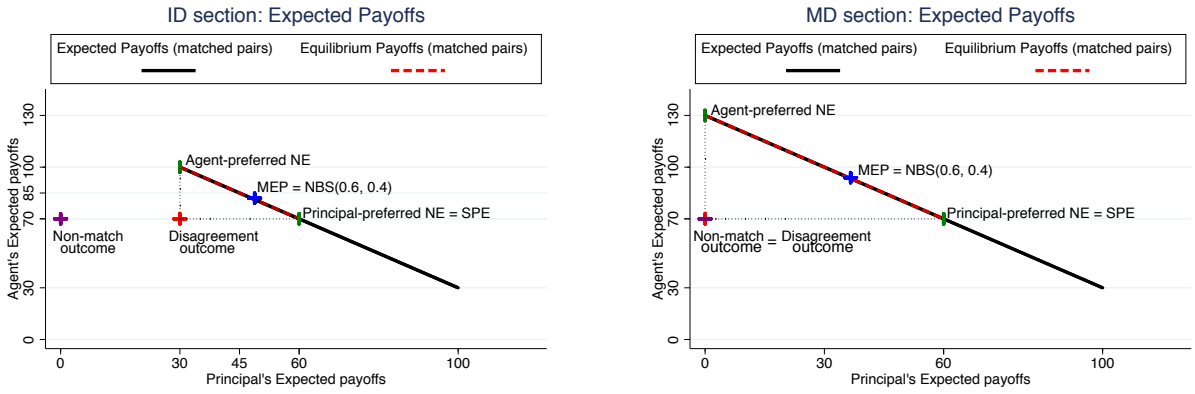


Figure 5: Possible expected payoffs in the ID and MD games, disagreement outcomes, average Matched Expected Payoffs (MEP), and Nash Bargaining Solution with bargaining weights 0.6 and 0.4.

To see if the behavior in both games can be explained by a 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 participants.[16] Blue crosses on Figure 5 show the overall average MEP. In Appendix A.4 we calculate which NBS bargaining power weights generate the MEPs in each game and find that they correspond to the principals' average bargaining weights of 0.593 and 0.610 in the ID and MD games respectively, a remarkably similar result (standard errors: 0.09 and 0.06). This provides strong evidence that disagreement outcomes are very important for the choices of the participants. Specifically, the fact

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

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

that principals have much more to lose than agents in case of disagreement leads to the increase of agents' bargaining power to 40% as compared to 0% predicted by the non-cooperative (PP)SPE.

Result 4. *Accounting for each player's minimum guaranteed expected payoffs, the Nash Bargaining Solution with bargaining weights (0.6, 0.4) predicts the average Matched Expected Payoffs in both ID and MD.*

4.4 Principals' Individual Behavior

In order to understand what causes the behavioral phenomena described in Sections 4.1 and 4.2, we analyze the determinants of the principals' success, as defined by their average payoffs in ID and MD, by looking at the individual choices and reactions to feedback. We start with the analysis that connects principals' average choices with their individual characteristics—fairness attitudes and the perception of rational cutoff points—elicited in Section 3 of the experiment. The regressions in Table 8 in Appendix C show that principals' average choices are not determined by their fairness considerations (variable Dictator choice).[17] Their perception of rational cutoff points (variable Cut-off) only matters slightly in ID, but not in MD (we discuss this in more detail in Section 5). This result supports our findings from Section 4.2 that principals adjust their choices to those of the agents due to the threat of not matching.

Section	Feedback State	$X \geq Y$	Ball	Agent's guess	Principal's payoff	Agent's payoff
ID	A	Yes	Blue	Blue	0	100
ID	B	Yes	Red	Red	100	100
ID	C	Yes	Blue	Red	100	0
ID	D	No	Red	Blue	0	0
ID	E	No	Blue	Blue	0	100
MD	B	Yes	Red	Red	$100 - X$	$100 + X$
MD	C	Yes	Blue	Red	$100 - X$	X
MD	D	No	Red	Blue	0	0
MD	E	No	Blue	Blue	0	100

Table 4: Principals' feedback states in the ID and MD games.

Before we get to the analysis of individual behavior, notice that there is a very strong correlation between principals' average Expected Payoffs in ID and MD (Spearman's $\rho = 0.73$, $p < 0.0001$). Specifically, some principals win a lot in both games and some win very little. Moreover, the independent groups of participants consist of a mixture of successful and unsuccessful principals (see Figure 12 in Appendix D). This suggests that success in both games is determined to a large extent by the individual choices that principals make and not by the groups that they are in, which allows us to concentrate on the individual choices of principals irrespectively of the group they belong to.

[17] The description of all variables use in the regressions can be found in Appendix B.

To understand the behavioral differences between the successful and the unsuccessful principals in the two games we divide them into high-earning and low-earning by the median of the sum of their Expected Payoffs in ID and MD. We consider the reactions of principals to "feedback states" that distinguish different types of situations that they face after each period. Table 4 shows the feedback states that depend on 1) whether the pair is matched ($X \geq Y$); 2) the color of the ball drawn and 3) the agent's final guess. Notice that the states unambiguously determine the payoffs of the players, however the opposite is not true: principals and agents can receive the same payoff in different states. Thus, our analysis includes not only the mechanistic reactions to the realized payoffs, which are only partially determined by the moves of Nature, but also the reactions to the outcomes of the principal-agent interaction.

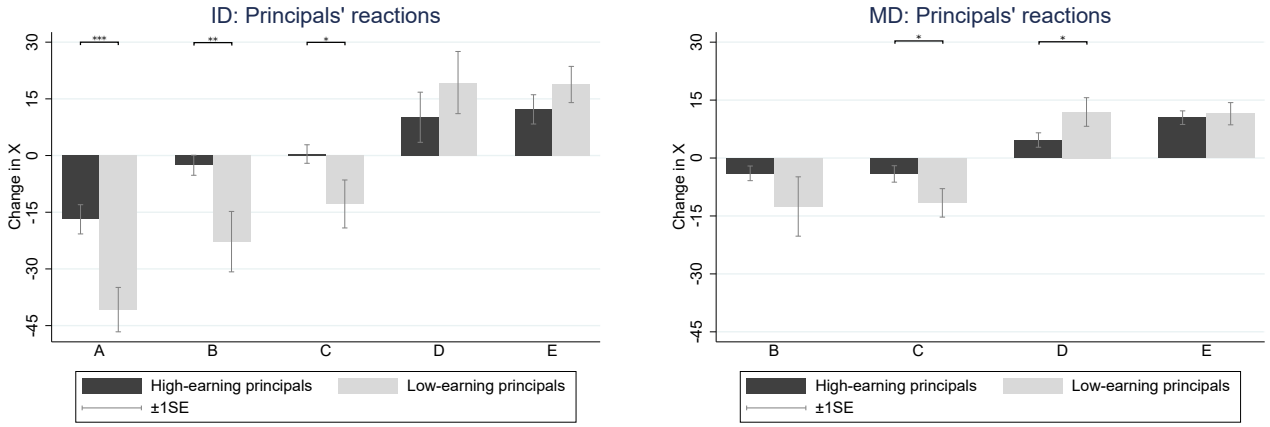


Figure 6: Principals' reactions to feedback states. The bars show the sums of coefficients from the random effects regressions reported in Table 9 in Appendix C. Significance levels *, **, *** correspond to $p < 0.1, 0.05, 0.01$.

Figure 6 shows the reactions of principals as estimated by the random effects regressions reported in Table 9 in Appendix C with changes in X as dependent variables and the dummies for the feedback states and low-earning principals as independent variables.[18] A simple observation from the graphs is that principals react to a match and a non-match in the expected direction: in both games they weakly decrease X after a match (states A, B, C) and increase X after a non-match (states D, E).[19] However, the sizes of these reactions are very different for high- and low-earning principals. Specifically, the more successful principals react to feedback in a much more reserved way than the less successful ones. This is especially pronounced in ID for the states in which there was a match (A, B, C). These observations suggest that the less successful principals overreact to the news that they successfully persuaded their paired agent by dramatically lowering X which in turn results in a significantly lower chance of matching in the next period. In addition, such reactions should lead to much more erratic behavior of the low-earning principals as compared to

[18] The regressions also control for the individual characteristics elicited in Section 3.

[19] We say "expected direction" in the sense that after a match (when $X \geq Y$) a principal should update her beliefs of the agents' actions downwards and thus lower X in an attempt to increase the payoff while still matching in the next period. Correspondingly, after a non-match ($X < Y$) a principal should update her beliefs of the agents' actions upwards and thus increase X in order to increase the probability of matching in the next period.

the high-earning ones. This is indeed the case. Table 5 shows that in ID low-earning principals have much higher standard deviations of choices X than high-earning principals (ranksum test, $p = 0.0001$). In MD this is also true, though the result is weaker (ranksum test, $p = 0.0537$).

Section	Measure	Sample	Principals	
			High-earning	Low-earning
ID	SD of Choices	All pairs	16.7 (2.10)	31.2 (2.17)
MD	SD of Choices	All pairs	13.2 (1.86)	19.3 (2.41)
N of observations			54	54

Table 5: Mean standard deviations of principals' choices X. Numbers in brackets indicate standard errors.

Result 5. *Principals who are more stable in their choices and react less to irrelevant circumstantial feedback earn significantly more than principals who are highly reactive to feedback and thus exhibit erratic behavior. This is true in both games, but is especially strong in ID.*

4.5 Agents' Individual Behavior

The individual choices of agents clearly have a significant influence on principals' earnings in both games. Thus in this section we analyze what drives their choices. As with principals, we first look at the connection between agents' average choices and their individual characteristics elicited in Section 3. The regressions in Table 8 in Appendix C show that fairness considerations are a significant determinant of agents' average choices in both games (variable Dictator choice). Generous agents who choose low Y in the Dictator tasks of Section 3 also choose lower average Y in the corresponding game. The effect is very strong in ID, in size and statistical significance, and somewhat smaller in size but still significant in MD. This is consistent with our previous findings that agents' choices are stable across the two games and are not influenced much by principals' actions. The rational cutoff perceptions influence agents' decisions only weakly in MD (10% significance level), and not at all in ID (we return to the discussion of cutoff points in Section 5).

Since fairness considerations seem to matter for agents' choices we consider the individual reactions to feedback states separately for "selfish" and "generous" agents, whom we define by the median split of their answers in the Dictator tasks.[20] Table 6 shows selfish and generous agents' average choices in the two games. The differences between the two groups in each game are large and significant (ranksum tests, ID: $p = 0.0001$; MD: $p = 0.0002$). In addition to the above mentioned regressions in Table 8, this supports our conclusion that fairness considerations have a large effect on agents' choices.

[20]We divide agents into selfish and generous separately for ID and MD since we find only weak correlation in ID and MD generosity within subjects.

Section	Measure	Sample	Agents	
			Selfish	Generous
ID	Choices	All pairs	70.6	50.0
			(2.43)	(3.99)
MD	Choices	All pairs	66.1	48.5
			(2.90)	(3.13)
N of observations			54	54

Table 6: Average choices of selfish and generous agents. Standard errors in parentheses.

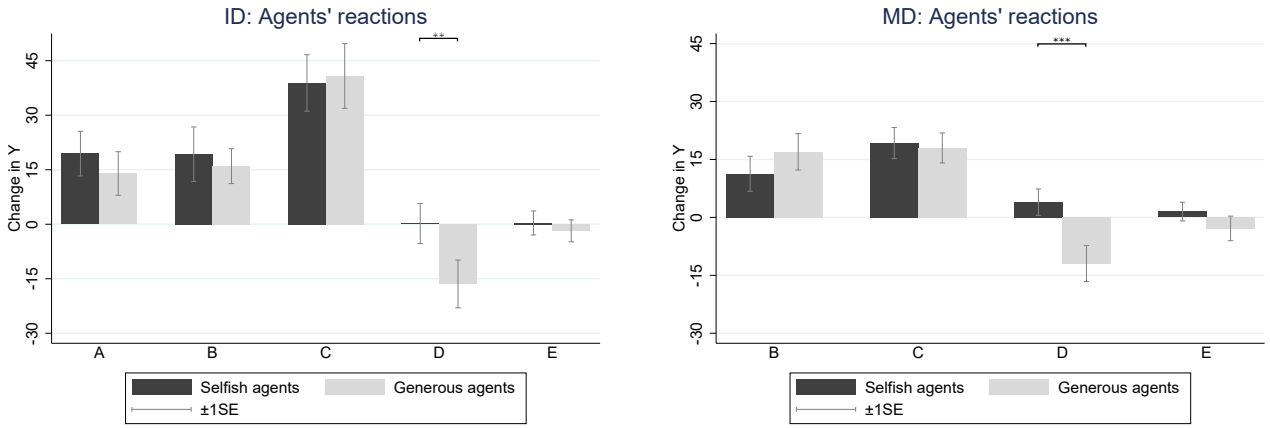


Figure 7: Agents' reactions to feedback states. The bars show the sums of coefficients from the random effects regressions reported in Table 9 in Appendix C.

Figure 7 shows agents' reactions depending on the feedback states as estimated by the random effects regressions reported in Table 9 in Appendix C.[21] We observe that agents' reactions are the opposite to the principals' reactions shown in Figure 6. After a match (states A, B, and C) agents increase Y and after a non-match (states D and E) they weakly decrease Y. This is consistent with our idea that participants treat the two games as a form of bargaining. The only significant difference between selfish and generous agents is in state D where there is no match and agents receive 0 points. Generous agents tend to significantly decrease Y, whereas selfish agents keep it at the same level. The fact that generous agents prefer lower Y than selfish ones in the Dictator tasks suggests that this decrease might be due to their fairness considerations. This demonstrates the mechanism responsible for the lower choices of generous agents as was reported in Table 6 above.

Result 6. *Agents' choices are strongly influenced by their fairness considerations, which in turn affect the level of principals' success in both games. Generous agents make an effort to match with principals by decreasing Y after a non-match, selfish agents do not.*

[21] The regressions also control for the individual characteristics elicited in Section 3 and the lagged choice of the principals that agents observe on their feedback screens.

5 Discussion

Despite the parallelism of information and mechanism design discussed in the literature and the theoretical equivalence of the games in our experiment, we find that the two environments are perceived very differently, and as a consequence the behavior in ID and MD is not the same and departs systematically from the theoretical predictions. In this section we discuss possible reasons why theoretical parallelism fails behaviorally.

ID and MD Entail Different Degrees of Payoff Uncertainty. From the theoretical perspective a risk neutral decision maker is indifferent between a lottery and its certainty equivalent. In some sense, this indifference lies at the core of the parallelism between information and mechanism design suggested in the literature. Nevertheless, the two designs are very different in terms of the payoff uncertainty experienced by the players. In the ID game players either win the highest possible payoff or nothing, whereas in the MD game the "deterministic" monetary transfers smooth out these extreme payoffs thus making the uncertainty created by the moves of Nature much less pronounced.

This difference between the two designs has major implications for the behavior of the participants in our experiment. Looking at Figures 6 and 7 it becomes clear that the reactions of principals and agents to the feedback are much larger in absolute values in ID than in MD, and, as we discussed above, this has serious ramifications for the earnings of the principals. One possible reason why the reactions are so different is that in ID participants always receive 0 points or 100 points, while in MD the payoffs are smoothed by the transfers. Provided that people are reinforcement learners, it is just natural that they overreact to extreme payoffs. Notice that this has little to do with possible risk aversion of the participants, but rather with the extreme nature of the payoffs in ID and the much more continuous payoffs in MD. Thus, while equivalent in expectation, the extreme nature of payoffs in ID requires a higher degree of expectations-based reasoning and self-restraint. This constitutes a major practical hurdle for the implementation of information design and its substitutability with mechanism design.

ID and MD Impose Different Kinds of Contracts. Another stumbling block to the theoretical parallelism of ID and MD is the difference in the "flavor" of a contract between the principal and the agent in the two games. The situation in MD is more straightforward and in some sense represents a more natural agreement: the principal pays money to the agent and in return the agent chooses an unfavorable option (guesses red). Even if the ball turns out to be blue and the agent does not earn any money from the guess, she is not particularly disappointed since she is compensated for exactly this contingency by the principal. This is reflected in Figure 7 where agents' reactions to different ball colors in case of a match (right graph, feedback states B and C) are very similar, even though the realized payoffs differ by 100 points. In ID, however, the situation is very different. When the agent agrees to follow the principal's recommendation there is a sense in which the agent

is "entrusting" the principal with making the guess. Thus, when it happens that the followed recommendation is bad (the guess is red but the ball is blue) the agent is left with no monetary compensation; a feeling of regret for not going with the default choice (guess blue); and a feeling of betrayal since the principal talked her into guessing a wrong color. Thus, there is a large room for discontent and resentment, which is clearly visible on the left graph of Figure 7 (feedback state C), which shows that when this happens agents tend to react very strongly by increasing Y in the next period by around 40 points. In fact, an additional analysis in Appendix E demonstrates that experiencing C in the ID game has a long-lasting impact on agents' behavior since their choices never return back to the pre-state C levels.

Overall, discontinuous payoffs and the necessity to fully trust principals' recommendations in information design seem to generate inflated emotion-driven reactions that result in behavior that is very different from what is happening in mechanism design characterized by continuous payoffs and pre-agreed compensations for bad outcomes. These differences should be considered by a principal who is choosing between designing an informational or a monetary contract.

Perception of Monetary and Informational Compensations. The difference in the perception of the ID and MD games does not end with the effects described above. Our data reveal that agents are more demanding when being incentivized than when being persuaded. However, if we take into account agents' estimates of the rational cutoff points the picture becomes somewhat different.

Figure 8 shows the average choices of principals and agents in ID and MD as well as their average estimates of the rational risk neutral cutoffs that are elicited by the Cutoff tasks in Section 3. Consider first the graph for MD. Here we see that both principals' and agents' average choices are close to what they think is the optimal cutoff. Thus, despite the fact that these estimates are above the theoretical predictions, agents ask for what they believe to be optimal on average. This is also supported by the regression in Table 8 in Appendix C, which shows that the agents' average choices in MD depend significantly, albeit at 10% level, on the expressed cutoff point.

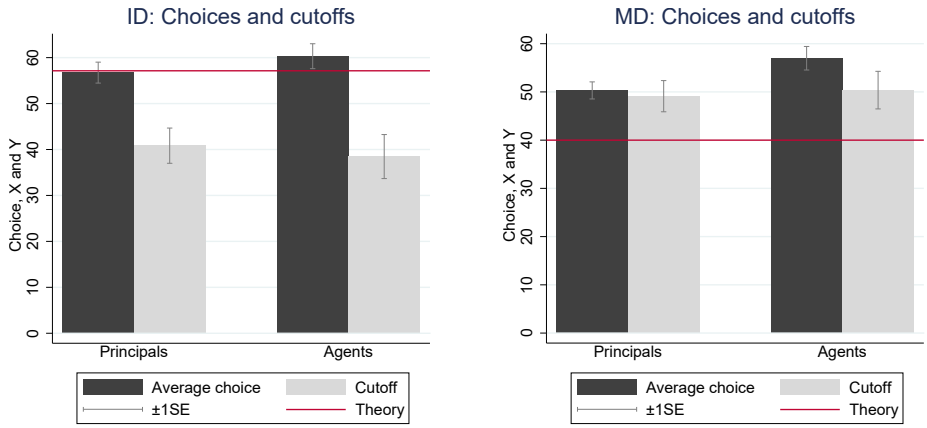


Figure 8: Average choices, theoretical predictions, and participants' estimates of rational (risk neutral) cutoffs.

In ID, however, the situation is very different. First, participants' average choices are very far from what they believe is optimal. Second, even though the average choices are close to the theoretical predictions, the much lower cutoff estimates suggest that this is coincidental. We hypothesize that, given the high uncertainty related to two independent sources—the moves of Nature and the probabilistic signals from principals—agents try to be more cautious and increase their choices above the level that they think is optimal, which also drags principals' choices along. This idea is not that far fetched. Indeed, many studies (e.g., Kahneman and Tversky, 1973; Charness *et al.*, 2007; Holt and Smith, 2009; Abdellaoui *et al.*, 2015) show that people are not very good at maximizing expected payoffs when there is a need to reason about probabilities or perform Bayesian updating. This should act as a warning to both policymakers and designers of informational environments. Receivers of information may not be particularly good at aggregating information and deciding how much of it is "enough." What is more, they seem to anticipate their limitations and consequently behave in a conservative manner in order to hedge against potential mistakes.

Theoretical Implications. Our final note is on the theoretical implications of our behavioral findings. As we discuss in Section 4.3, the Principal-Preferred Subgame-Perfect Equilibrium concept of Kamenica and Gentzkow (2011) does not seem to do a good job at explaining the behavior of participants since it completely ignores the bargaining power that agents have in both the ID and the MD games. Of course, Kamenica and Gentzkow (2011) consider a one-shot interaction, whereas our participants play many games in a row which allows principals to get an idea about the agents' *strategy*, the threshold value of Y. This is exactly where the bargaining power may become important. However, as is well known from, for example, Ultimatum bargaining (Güth *et al.*, 1982), even in one-shot interactions the ability of the second mover to destroy the payoff of the first has a very tangible effect on the first movers' choices. Notice as well that this effect should be more pronounced in the ID and MD games than in the Ultimatum game, since in the latter the second mover destroys her own payoff along with the first mover's (and first movers still choose to accommodate), while in our games the agent guarantees herself a rather high expected payoff of 70 points by not following the recommendation or rejecting the transfer. Thus, the threat to the principal is much more severe. We think that future models of Information Design should take this very real effect into account if they strive to explain human behavior in ID and MD environments.

6 Conclusion

This paper presents an experimental analysis of the theoretical parallelism between information design (ID) and mechanism design (MD) that has been proposed in the literature (see e.g., Bergemann and Morris, 2017; Taneva, 2017). We modify the framework of Kamenica and Gentzkow (2011) so that both the ID and MD principal-agent problems can be formulated 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].

In order to minimize various contextual influences we have designed the ID and MD games in such a way that both principals and agents choose a single number 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). An agent receives a reward if she guesses the color of the ball correctly, while a principal receives a reward if the agent guesses color red. In MD a principal chooses the amount of money to pay an agent for choosing red. In ID a principal chooses the informativeness of a message sent to an agent about the color of the ball. The similarity of the two games allows us to make direct comparisons of benefits of persuasion and incentivization.

We find that the (PP)SPE is a poor predictor of actual choices in both games. One reason is that agents are much more demanding when the problem is framed as a mechanism design game than when it is framed as an information design game. Another reason is that in (PP)SPE principals have a serious first mover's advantage, which should theoretically allow them to extract all surplus from the agent, whereas in reality principals are threatened by the possibility of disagreement with the agent in which case they receive zero payoff, while agents still receive high profits from guessing blue. To stress the behavioral importance of the disagreement outcomes we analyze the games with the Nash Bargaining Solution and find that agents possess bargaining power of about 40%, which explains average payoffs in both games.

Our main result is that ID is more profitable for principals than MD. Specifically, principals in ID are more often successful at persuading agents to follow their recommendation than to accept a transfer in MD. Moreover, principals earn significantly more in ID than in MD conditional on a successful persuasion or incentivization. Additionally, we find that in ID, which admits only extreme payoffs (win all or lose all), some principals overreact to irrelevant information about the ball color, which leads to their earning much less than the principals who ignore this information. In MD this problem does not arise since the payoffs are smoothed by the transfers. Agents also react differently to similar outcomes in the two games. In ID, when agents follow the recommendation but fail to guess the correct color, their reaction is much more extreme than in MD when they agree to a transfer, guess red, but the ball is blue. We hypothesize that this difference comes from "betrayal aversion" experienced in ID but not in MD. Finally, we find that agents have more difficulties in understanding how to maximize their payoffs in ID as compared to MD, which results in more reliance on simple heuristics like "fair" outcomes and noisier choices that give advantage to principals.

Overall, while theoretical parallels can be easily drawn between the emerging field of information design and the more traditional mechanism design, our study shows that in practice this parallelism faces strong challenges. The main reason is that people perceive informational manipulations differently from monetary ones. Our study singles out some of the behavioral aspects of

this difference, but further investigations are necessary to fully comprehend the mechanisms that determine behavior in the two environments.

References

Abdellaoui, M., Klibanoff, P. and Placido, L. (2015). Experiments on compound risk in relation to simple risk and to ambiguity. *Management Science*, **61** (6), 1306–1322.

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.

Au, P. H. and Li, K. K. (2018). Bayesian persuasion and reciprocity: Theory and experiment.

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.

Charness, G., Karni, E. and Levin, D. (2007). Individual and group decision making under risk: An experimental study of bayesian updating and violations of first-order stochastic dominance. *Journal of Risk and uncertainty*, **35** (2), 129–148.

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., Lizzeri, A. and Perego, J. (2018). *Rules and commitment in communication*. Tech. rep.

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.

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

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

Holt, C. A. and Smith, A. M. (2009). An update on bayesian updating. *Journal of Economic Behavior & Organization*, **69** (2), 125–134.

Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological review*, **80** (4), 237.

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. (2017). On bayesian persuasion with multiple senders.

Mathevet, L., Perego, J. and Taneva, I. (2017). On information design in games.

Nguyen, Q. (2016). Bayesian persuasion: Evidence from the laboratory.

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]

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

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:

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

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

The intersection of the above best response correspondences identifies a continuum of Nash equilibria for each game as shown in Figure 9.

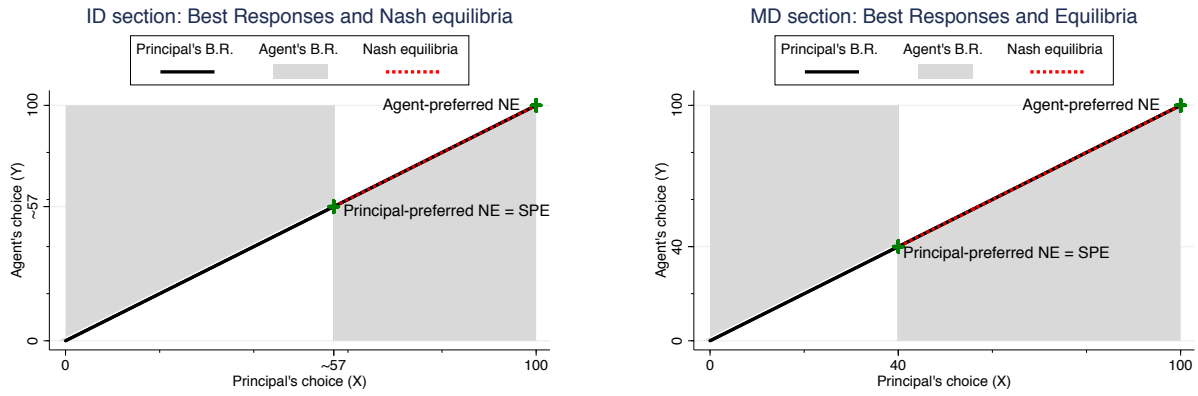


Figure 9: Best responses and Nash equilibria for the ID and MD games. "Agent-preferred NE" refers to the Nash equilibrium that is best for the agent in terms of expected payoffs. Correspondingly "Principal-preferred NE" refers to the Nash equilibrium that is best for the principal.

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 10 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:

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

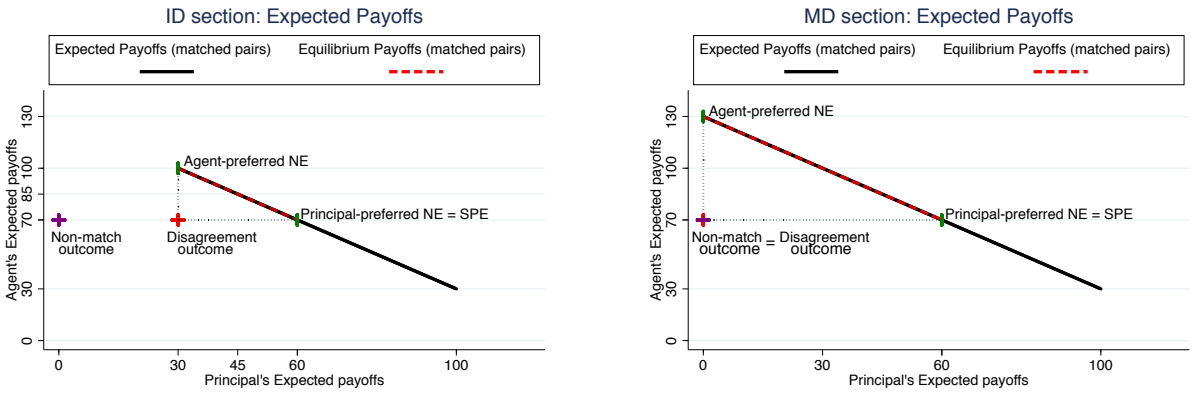


Figure 10: Expected payoffs in the ID and MD games. "Disagreement outcome" refers to the minimum guaranteed expected payoffs for each player. "Non-match outcome" refers to the expected payoffs when the principal fails to persuade/incentivize the agent ($X < Y$).

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 α 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 α and the agent's matched average expected payoff in the

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

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

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):

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

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 Variables Used in the Regressions

Variable	Range	Definition
Average choice	$[0, 100]$	for each participant and each game (ID or MD), the average choice (X or Y) made in 10 periods
Dictator choice	$[0, 100]$	for each participant and each game (ID or MD), the choice (X if principal, Y if agent) made in the Dictator tasks in Section 3
Cutoff	$[0, 100]$	for each participant and each game (ID or MD), the choice (X if principal, Y if agent) made in the Cutoff tasks in Section 3
D.X	$[0, 100]$	the difference in principal's choice X in the current and the previous period
D.Y	$[0, 100]$	the difference in agent's choice Y in the current and the previous period
Low-earning principal	$0/1$	is 1 if the average expected payoff of a principal in a game (ID or MD) is below median of all 54 principals
Generous agent in ID	$0/1$	is 1 if agent's cutoff estimate in ID (the Cutoff ID task in Section 3) is below median of all 54 agents
Generous agent in MD	$0/1$	is 1 if agent's cutoff estimate in MD (the Cutoff MD task in Section 3) is below median of all 54 agents
L.B, L.C, L.D, L.E	$0/1$	is 1 if the feedback state B, C, D, E was observed by a participant in the previous period
L.X in ID	$[0, 100]$	the principal's choice X in ID observed by an agent in the previous period
L.X in MD	$[0, 100]$	the principal's choice X in MD observed by an agent in the previous period

Table 7: Variables used in the regressions.

C Additional Regressions

Average choice	Principals		Agents	
	ID	MD	ID	MD
Dictator choice	–0.026	0.073	0.445***	0.215**
	(0.060)	(0.117)	(0.078)	(0.106)
Cutoff	0.186**	–0.059	0.036	0.161*
	(0.079)	(0.043)	(0.062)	(0.086)
Constant	49.483***	52.286***	27.776***	33.470***
	(3.566)	(2.237)	(8.966)	(7.333)
N observations	54	54	54	54
N groups	13	13	13	13

Table 8: Random effects linear regressions of average choices on the Dictator and Cutoff choices in Section 3. Errors are robust and clustered by the 13 independent groups of participants. Standard errors in parentheses. Significance levels *, **, *** correspond to $p < 0.1, 0.05, 0.01$.

D.X, D.Y	Principals		Agents	
	ID	MD	ID	MD
Low-earning principal	–23.896***	–8.585		
	(6.842)	(7.677)		
Generous agent in ID			–5.476	
			(4.606)	
Generous agent in MD				5.697
				(3.888)
L.B	14.305***		–0.168	
	(3.210)		(7.923)	
L.C	17.272***	–0.155	19.449***	7.963***
	(3.713)	(2.696)	(6.901)	(1.937)
L.D	27.005***	8.624***	–19.214***	–7.346*
	(7.379)	(2.581)	(7.250)	(3.920)
L.E	29.086***	14.414***	–19.083***	–9.773**
	(5.615)	(2.467)	(5.397)	(3.946)
L.B × Group	3.683		2.203	
	(8.380)		(8.955)	
L.C × Group	10.682	1.094	7.393	–6.972**
	(9.272)	(7.516)	(10.174)	(3.492)
L.D × Group	33.073**	15.841*	–11.170	–21.582***
	(13.560)	(8.753)	(10.405)	(7.437)
L.E × Group	30.477***	9.606	3.309	–10.073*
	(10.125)	(9.958)	(7.064)	(5.789)
Dictator choice	–0.036	0.013		
	(0.031)	(0.022)		
Cutoff	0.064	–0.011	–0.018	0.054**
	(0.050)	(0.018)	(0.029)	(0.023)
L.X in ID			–0.145***	
			(0.055)	
L.X in MD				–0.179***
				(0.044)
Constant	–16.873***	–3.964**	19.435***	11.293**
	(3.864)	(1.899)	(6.145)	(4.527)
N observations	486	486	486	486
N groups	54	54	54	54

Table 9: Random effects linear regressions of change in choice (X or Y) between periods t and $t-1$. The dummy Group stands for Low-earning principal, Generous agent in ID, or Generous agent in MD depending on the regression. Errors are robust and clustered by participant. Significance levels *, **, *** correspond to $p < 0.1, 0.05, 0.01$.

D Additional Figures

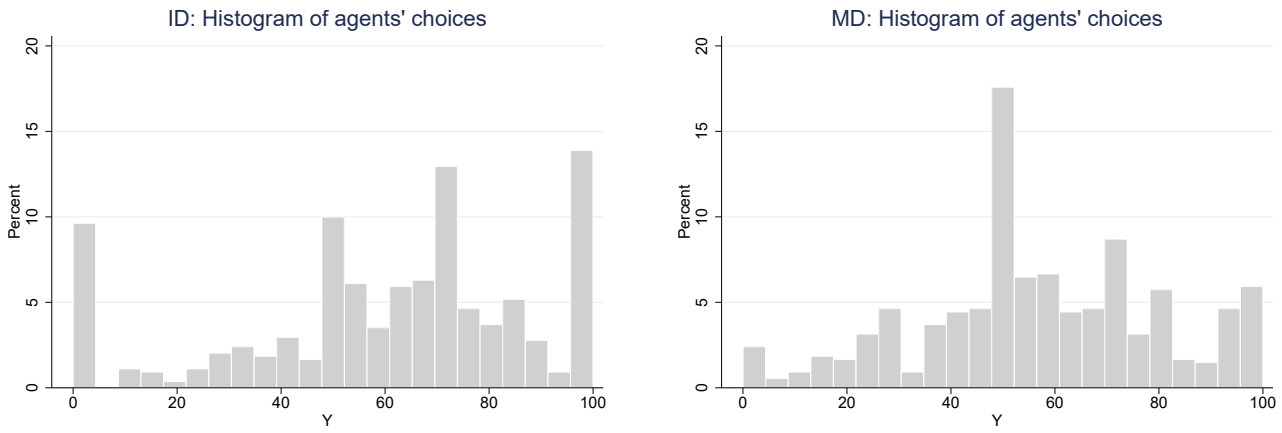


Figure 11: Histograms of agents' choices in ID and MD.

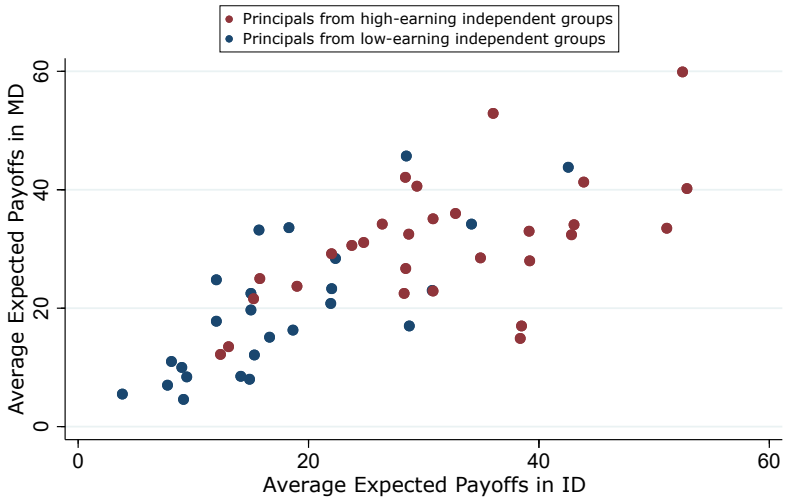
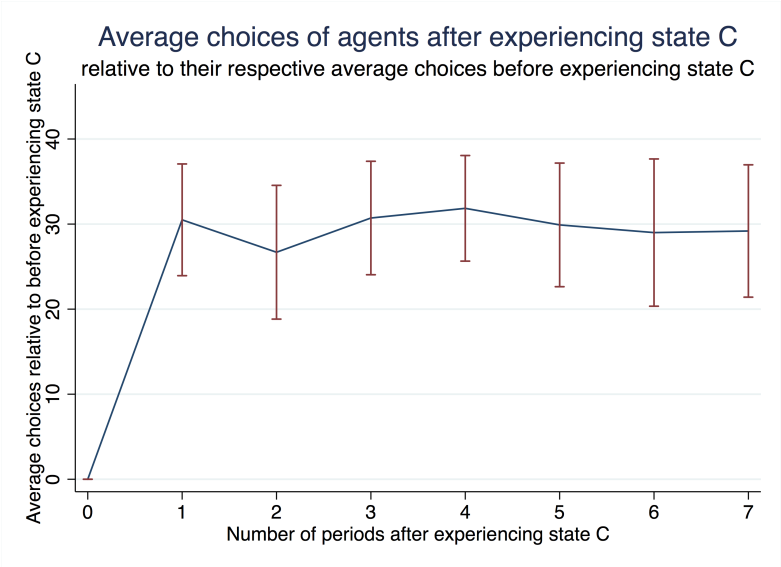


Figure 12: Scatterplot of the principals' average Expected Payoffs (54 principals). We divide the 13 independent groups of principals into high-earning and low-earning groups by the median split of average group payoffs. The red dots show principals from the high-earning groups and the blue dots from the low-earning groups. It is clear that some principals from high-earning groups earn little and vice versa. Therefore, earnings are not exclusively determined by the group to which a principal belongs, but also depend on her individual decisions.

E The Impact of Feedback State C in ID

In the analysis of the reactions of agents to feedback states in Section 4.5 we observed that the largest reaction in ID was to state C, which happens when agents follow the recommendation, guess red, but receive 0 points because the ball is blue. As we discuss in Section 5 this reaction is the main difference between ID and MD since in MD agents are paid for guessing red, which does not create the "betrayal" problem evident in ID. We hypothesize that agents blame principals for "tricking" them into following the recommendation.



The figure above shows the average increase in choices of agents after experiencing feedback state C for the first time. For each agent we take the difference between the consecutive choices and the choice that was made when state C happened. We observe that agents are influenced by state C so much that they do not decrease their choices even 7 periods after the occurrence of state C, indicating that experiencing this state does not only cause extreme but also long-lasting reaction.

F Instructions

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

F.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 **25 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 25 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").

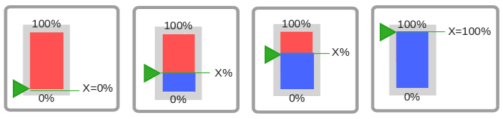
Player A will choose the % chance [X%] of giving the correct recommendation to Player B if the ball is BLUE (i.e. "Guess blue")

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").

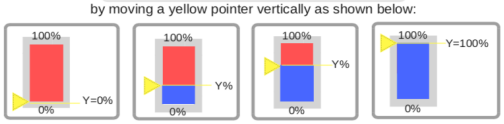
Player B will choose the minimum % chance [Y%] of receiving the correct recommendation when the ball is BLUE that Player B is willing to accept in order to follow Player A's recommendations.

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:

Recommendation	**Choices** of Player A & Player B: X ≥ Y	X < Y
"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)

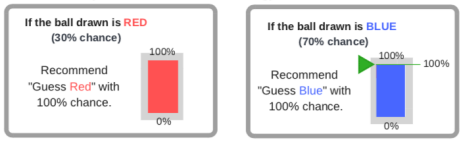
The rightmost example shows the case where a Player B is only willing follow Player A's recommendations if **recommendations are always correct** when the ball is **BLUE**. (since X needs to equal 100 in order to satisfy Y)

Examples: How are Player A's recommendations generated.

Suppose that **Player A's choice is 100**: If the ball is **BLUE**, the recommendation will **always** be "Guess Blue" (with 100% probability) and **never** "Guess Red" (with 0% probability).

Remember that if the ball is **RED**, the recommendation will always be "Guess Red".

Player A's **recommendation strategy** is as follows:



If the recommendation is "Guess Red", then the ball must be **RED** with certainty.
If the recommendation is "Guess Blue", then the ball must be **BLUE** with certainty.
By following this recommendation plan, Player B is guaranteed to guess the correct color of the ball and thus receive 100 points for sure.

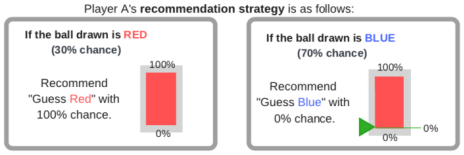
If Player B follows this recommendation plan, and since Player A gets 100 points only when Player B's guess is Red, Player A will receiver 100 points with 30% chance (if the ball drawn is **RED**).

If Player B does not follow this recommendation plan, Player B's color guess will be Blue. Since a **BLUE** ball is drawn with 70% chance, Player B will receive 100 points with 70% chance. Player A will receive 0 points for sure.

Suppose that **Player A's choice is 0**: If the ball is **BLUE**, the recommendation will **never** be "Guess Blue" (with 0% probability) and **always** "Guess Red" (with 100% probability).

Remember that if the ball is **RED**, the recommendation will always be "Guess Red".

Player A's **recommendation strategy** is as follows:



The recommendation plan will always recommend "Guess Red" regardless of the color of the ball.

If Player B follows this recommendation plan, Player B's color guess will be Red with certainty. Since a **RED** ball is drawn with 30% chance, Player B will receive 100 points with 30% chance. Player A will receive 100 points for sure since Player B always guesses Red.

If Player B does not follow this recommendation plan, Player B's color guess will be Blue. Since a **BLUE** ball is drawn with 70% chance, Player B will receive 100 points with 70% chance. Player A will receive 0 points for sure.

Summary of the procedure in each round

1

The computer randomly draws a ball from an urn containing 3 **RED** balls and 7 **BLUE** balls.

2

Players make their choices simultaneously

Player A chooses (X): the % chance of giving correct recommendation if the ball is **BLUE**

Player B chooses (Y): the minimum % chance of receiving correct recommendation if the ball is **BLUE**, that Player B is willing to accept in order to follow Player A's recommendations

3

Outcomes and earnings are determined

A recommendation is generated according to Player A's choice of % of correct recommendations (X) if the ball is **BLUE**.

If X $\geq$ Y, Player B will follow Player A's recommendations.
Player B's color guess will then automatically be:
Red if the generated recommendation is "Guess red" and
Blue if the generated recommendation is "Guess blue".

If X < Y, Player B will ignore Player A's recommendations.
Player B's color guess will then automatically be Blue, regardless of the recommendation.

Both players are told the color of the ball, the recommendation generated, whether Player B followed or ignored Player A's recommendation, Player A's choice (X) and Player B's color guess. **Player A will not be told Player B's choice of Y.**

Each Player learns their respective earnings.

F.2 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 **25 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 25 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.

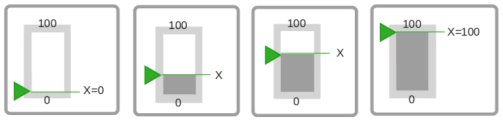
Player A will choose the number of points [X] to transfer to Player B if Player B guesses Red.

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.

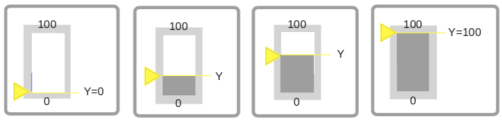
Player B will choose the minimum transfer of points [Y] from Player A that Player B is willing to accept in order to guess Red.

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)

Summary of the procedure in each round

1

The computer randomly draws a ball from an urn containing 3 **RED** balls and 7 **BLUE** balls.

2

Players make their choices simultaneously

Player A chooses (X): the number of points to transfer to Player B if Player B guesses Red.

Player B chooses (Y): the minimum transfer of points from Player A that Player B is willing to accept in order to guess Red.

3

Outcomes and earnings are determined

If X $\geq$Y, Player B will accept Player A's proposed transfer.
Player B's color guess will then automatically be Red.
Player A will transfer (X) number of points to Player B regardless of the ball color.

If X < Y, Player B will reject Player A's proposed transfer.
Player B's color guess will then automatically be Blue.
No transfer will take place between the players.

Both players are told the color of the ball, whether Player B accepted or rejected Player A's recommendation, Player A's choice (X) and Player B's color guess.
Player A will not be told Player B's choice of Y.

Each Player is told their respective earnings.

G Screenshots

G.1 ID Section

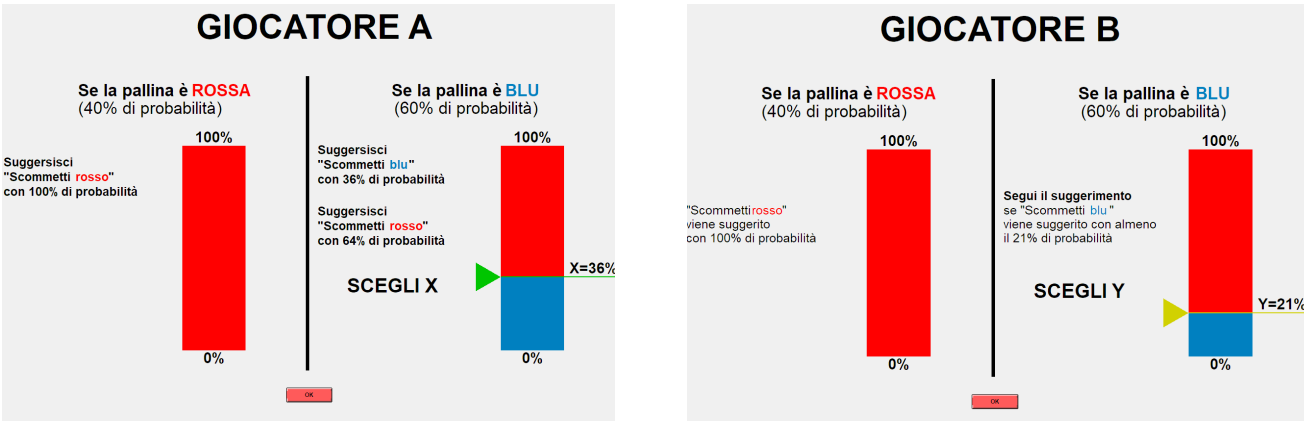


Figure 13: Choice screens of Player As - principals ("Giocatore A") and Player Bs - agents ("Giocatore B") in each period in the ID section.

Figure 14: Feedback screens of Player As - principals (Left) and Player Bs - agents (Right) in each period in the ID section.

G.2 MD Section

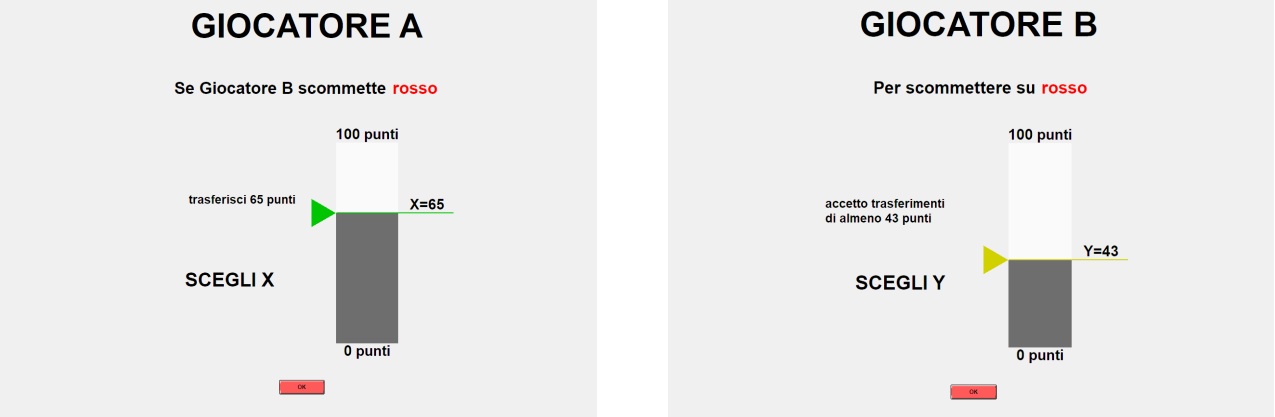


Figure 15: Choice screens of Player As - principals ("Giocatore A") and Player Bs - agents ("Giocatore B") in each period in the MD section.

Figure 16: Feedback screens of Player As - principals (Left) and Player Bs - agents (Right) in each period in the MD section.