Persuasion Bias in Science: An Experiment

Arianna Degan (University of Quebec at Montréal, CIRPÉE)
Ming Li (Concordia University, CIREQ, CIRANO)
Huan Xie (Concordia University, CIREQ, CIRANO)

Workshop in Memory of Artyom Shneyerov
October 12, 2018

Motivations

- Can we use economics models (game theoretical models) to examine incentives and welfare in research conduction?

Motivations

- Can we use economics models (game theoretical models) to examine incentives and welfare in research conduction?
- Specifically, we investigate a situation that applies persuasion to scientific research.
  - Conflicts of interests between Researcher and Evaluator
  - Asymmetric information between Researcher and Evaluator
  - Researcher tries to persuade Evaluator the existence of positive treatment effect

Motivations

- Can we use economics models (game theoretical models) to examine incentives and welfare in research conduction?
- Specifically, we investigate a situation that applies persuasion to scientific research.
    - Conflicts of interests between Researcher and Evaluator
    - Asymmetric information between Researcher and Evaluator
    - Researcher tries to persuade Evaluator the existence of positive treatment effect
- Examples: pharmacy industry, publishing papers, applying for grants

Questions

Game theoretical model not replying on reputation or social preference

- Do researchers have incentives to cheat?
- Can evaluators predict the bias and correct their evaluation accordingly?
- Impact on welfare

Literature

- The project is related to the broad literature on communication and information transformation (Crawford and Sobel, 1982), especially the arising literature on persuasion (Kamenica and Gentzkow, 2011).
    - Blume, Lai and Lim (2017): Survey of experiments and theoretical foundations on strategic information transmission
    - Experimental studies on persuasion game: Frechette, Lizzeri, and Perego (2017), Nguyen (2017), which focus on the effect of commitment.

Literature

- The project is related to the broad literature on communication and information transformation (Crawford and Sobel, 1982), especially the arising literature on persuasion (Kamenica and Gentzkow, 2011).
  - Blume, Lai and Lim (2017): Survey of experiments and theoretical foundations on strategic information transmission
  - Experimental studies on persuasion game: Frechette, Lizzeri, and Perego (2017), Nguyen (2017), which focus on the effect of commitment.
- Theoretical studies on scientific research
  - Di Tillio, Ottaviani and Sørensen (2017a, 2017b)
  - Our experiment is based on a simplified model of Selective Sampling in Di Tillio, Ottaviani and Sørensen (2017a)

Model: Di Tillio, Ottaviani and Sørensen (2017a)

- Use a game-theoretical framework to model randomized controlled trial (RCT)
- Three cases of possible manipulation by researchers
  - **Selective sampling**: non-randomly select sample $\Rightarrow$ undermine the external validity
  - Selective assignment: non-randomly assign subjects into treatment $\Rightarrow$ undermine the internal validity
  - Selective reporting $\Rightarrow$ challenge both internal and external validity

Model: Basic Elements

- Two risk-neutral players: Researcher and Evaluator
- Researcher sets up an experiment.
- Evaluator observes the experiment outcome and decides whether to grant Researcher a desired acceptance (e.g., a funding award or a journal publication).
- The aim of the experiment is to estimate the effect of a treatment (e.g., by a new drug or a new policy).
- Evaluator only grants acceptance if the average treatment effect is strong enough compared to the cost of acceptance k.
- Researcher always benefits from acceptance.

Model: Treatment Effects

- The experiment can be conducted in one of two locations: Left or Right.
- Population is equally divided between the two locations.
- For simplicity, assume all individuals in one location have the same treatment effect: $\beta_L, \beta_R \in \{0, 1\}$
- β_L, β_R are i.i.d. across locations:
  $\Pr(\beta_L = 1) = \Pr(\beta_R = 1) = q$
  $\Pr(\beta_L = 0) = \Pr(\beta_R = 0) = 1 - q$
- Average Treatment Effect for the entire population:
  $\beta_{ATE} = (\beta_L + \beta_R)/2$

Model: Experiment Outcome/Evidence

- Location where the experiment is conducted: $t = L, R$
- Baseline experiment outcome: 0
- Experiment outcome under treatment conducted at location t: $v = \beta_t$
- From previous assumption β_L, β_R are i.i.d.
  - $\Pr(v = 1) = q$
  - $\Pr(v = 0) = 1 - q$
- Evaluator only observes the experiment outcome under treatment v, but not the location t where the experiment is conducted.
- $E(\beta_{ATE}|v)$: Evaluator's posterior expectation of the average treatment effect after observing experiment outcome v

Timing of the Game: No-manipulation

- Both players observe the Evaluator's cost of acceptance k.
- Researcher selects one location $t \in \{L, R\}$ to conduct the experiment.
- Evaluator chooses to accept or reject after observing the experiment outcome v.

Timing of the Game: Manipulation

- Both players observe the Evaluator's cost of acceptance k.
- **Researcher observes the true treatment effect in one location, β_A, $A \in \{L, R\}$.**
- Researcher selects one location $t \in \{L, R\}$ to conduct the experiment.
- Evaluator chooses to accept or reject after observing the experiment evidence v.

Researcher's Equilibrium Behavior

- No-manipulation: choose a location randomly
- Manipulation: Intuitive Strategy
  - If $\beta_A = 1$, choose $t = A$: If the private information reveals positive treatment effect, choose the location same as the one in the private information.
  - If $\beta_A = 0$, choose $t = -A$: If the private information reveals negative treatment effect, choose the location different from the one in the private information.

Effects of Manipulation

	No-manipulation		Manipulation	
	$E(\beta_{ATE}\|\cdot)$	w. p.	$E(\beta_{ATE}\|\cdot)$	w. p.
$v = 1$	0.75	0.5	0.67	0.75
$v = 0$	0.25	0.5	0	0.25

- Assume $\Pr(v = 1) = q = 0.5$: treatment effect is 1 with probability 0.5 and 0 with probability 0.5
- Manipulation increases the probability of positive experiment outcome
- Meanwhile, it decreases the conditional expectation of ATE, $E(\beta_{ATE}|\cdot)$
- Similar effects hold when $q \neq 0.5$

Evaluator's Equilibrium Behavior when $q = 1/2$

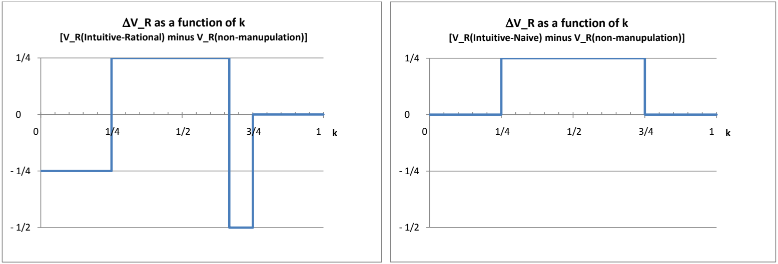
Evaluator's BR under No-manipulation

	$k \leq 0.25$	$0.25 < k \leq 0.75$	$k > 0.75$
$v = 1$	accept	accept	reject
$v = 0$	accept	reject	reject

Evaluator's BR under Manipulation

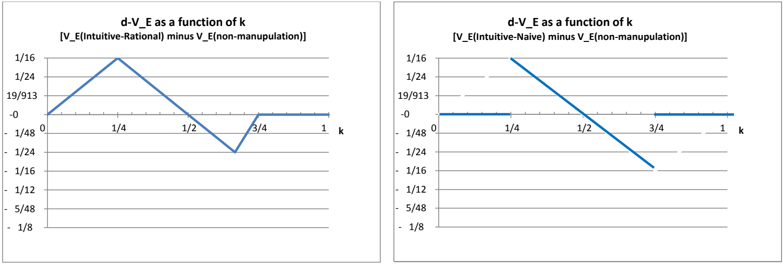
	$k \leq 0.67$	$k > 0.67$
$v = 1$	accept	reject
$v = 0$	reject	reject

Predictions on Welfare Analysis for Researcher



- Researcher's expected payoff under manipulation minus that under No-manipulation, as a function of k
- Left panel: rational Evaluator
- Right panel: naive Evaluator

Predictions on Welfare Analysis for Evaluator



- Evaluator's expected payoff under manipulation minus that under No-manipulation, as a function of k
- Left panel: rational Evaluator
- Right panel: naive Evaluator

Parameterization

- The probability of positive treatment effect in each location: $q = 0.5$
- Under manipulation, the probability that Researcher observes private information from each location: $m = 0.5$
  - Evaluator is not informed of the experiment location $\Rightarrow$ The value of m does not affect players' decision.
  - The value of m is not explicitly told to subjects.
- Payoffs and cost of acceptance multiplied by 100
- k =10, or 40, or 70
  - In theory k is revealed to both Researcher and Evaluator.
  - We choose to test the theory given several fixed k values rather than drawing k from a distribution every round.

Parameterization (Cont'd)

- The values of k are chosen to satisfy the following predictions:

		$k_1 = 10$	$k_2 = 40$	$k_3 = 70$
$v = 1$	Manipulation	accept	accept	**reject**
	No-Manipulation	accept	accept	**accept**
$v = 0$	Manipulation	**reject**	reject	reject
	No-Manipulation	**accept**	reject	reject

- The predictions not only hold for risk-neutral Evaluators, but also hold for risk-aversive Evaluators who have CRRA utility function u^r with $r = 0.5$.

Experimental Design

- Treatments: No-manipulation vs. Manipulation, different *k* value, Human Researcher vs. Robot Researcher
- Structure of a session

	No-Manipulation (30 rounds)			Manipulation (30 rounds)		
Cohort 1 of E	k1	k2	k3	k1	k2	k3
Cohort 2 of E	k2	k3	k1	k2	k3	k1
Cohort 3 of E	k3	k1	k2	k3	k1	k2

Experimental Design (Cont'd)

- We choose the order from No-manipulation to Manipulation for subjects to learn first in a simpler environment
- Instructions for Manipulation treatment only distributed upon the time to play
- Quiz after reading the instructions
- 3 practice rounds before each treatment starts

Experimental Design (Cont'd)

- Human Researcher treatment:
  - 12 subjects each session, 6 Researchers and 6 Evaluators, without changing player roles
  - Each round Researchers and Evaluators randomly and anonymously paired with each other. Researchers always face the same distribution of k.

Experimental Design (Cont'd)

- Human Researcher treatment:
  - 12 subjects each session, 6 Researchers and 6 Evaluators, without changing player roles
  - Each round Researchers and Evaluators randomly and anonymously paired with each other. Researchers always face the same distribution of k.
- Robot Researcher treatment:
  - Robot Researchers always follow the Intuitive Strategy.
  - Evaluators know the strategy used by Robot Researcher $\Rightarrow$ no strategy uncertainty
  - There is no interactions between Evaluators.

Implementation of the Game in a Round

Game environment:

- There are 50 balls in the Left Bin and 50 balls in the Right Bin.
- All balls in the same bin are of the same color.
- In each bin, the color of the balls is Red w.p. 50% and Blue w.p. 50%.
- Red balls have a value of 1 point and Blue balls have no value.

Implementation of the Game in a Round (Cont'd)

Game in the round:

- Both players observe k for the round. (k is described as Player B's endowed income.)
- If in the Manipulation treatment, Player A receives a private message about the color of the balls in one bin.
- Player A chooses one bin, Left or Right.
- The color of the balls in the chosen bin is shown to both players.
- Player B chooses whether to choose Implement the project.
  - If yes, Player B receives the value of the project, which equals the total number of red balls in the two bins, but has to give up the endowed income k. Player A receives 100 points.
  - If no, Player B receives k points. Player A receives nothing.

Payment

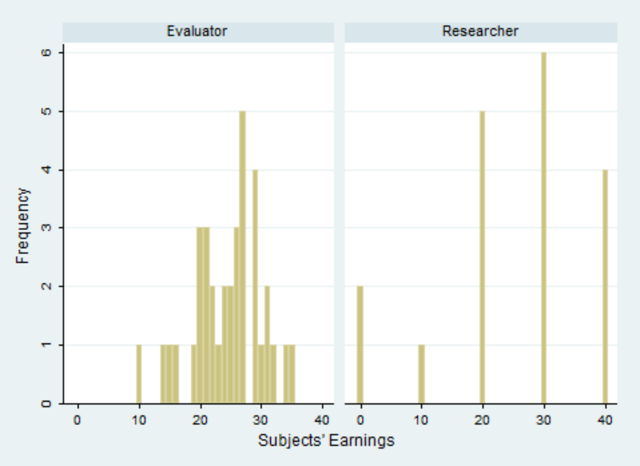
- At the end of the experiment, 2 rounds in each treatment are chosen for actual payment. In total, 4 rounds are paid.
- In every round, subjects are shown the history of play and previous payoffs from each round in that treatment.
- Points are converted to Canadian dollar at 10 points=$1.
- Show-up fee: $10
- If in the end subjects' total earning including show-up fee is less than $15, then they receive $15.

Sessions

- 3 sessions for Human Researcher treatment, with 18 pairs of Researchers and Evaluators
- 1 session for Robot Researcher treatment, with 18 Evaluators
- Treat each individual as an independent observation in conducting non-parametric tests
- Experiment conducted at CIRANO in Montreal, Canada

Earnings

Earning Distributions of Researchers and Evaluators

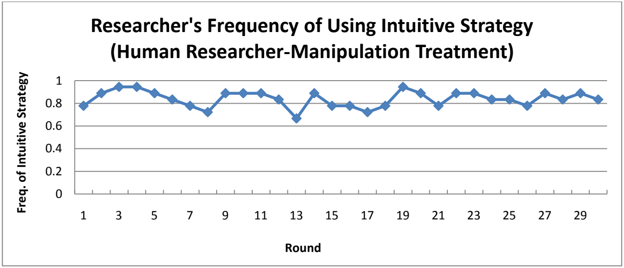


Earnings (Cont'd)

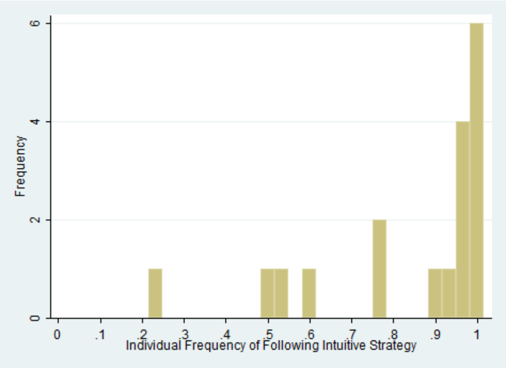
- Average earnings excluding show-up fee: \$25.19
- Researchers: Avg. \$25, Min \$0, Max \$40
- Evaluators: Avg. \$24.56, Min \$10, Max \$35
- No difference between Researchers' and Evaluators' earnings (Mann-Whitney test, $p = 0.51$)
- No difference in Evaluators' earnings between Human and Robot Researcher treatments (Mann-Whitney test, $p = 0.48$)

Researchers' Behavior

- Researchers' frequency of following the Intuitive Strategy in the Manipulation treatment
  - Avg. frequency 83.9%
  - The probability of adopting the Intuitive Strategy does not depend on the message content, k, or period.
  - No clear learning effect over time



Researchers' Ind. Freq. of Using Intuitive Strategy



Finding 1: *Researchers follow the Intuitive Strategy in the Manipulation treatment to a large extent.*

Evaluators' Behavior

Finding 2: Compared to the model prediction, Evaluators exhibit both significant over-implementation and under-implementation.

Evaluators' Behavior

Finding 2: Compared to the model prediction, Evaluators exhibit both significant over-implementation and under-implementation.

Finding 3: Overall the comparative statics are consistent with model predictions, especially in the Robot treatment.

Evaluators' Freq. of Implementation (Human Researcher)

No-manipulation (Part One)									
	$k = 10$			$k = 40$			$k = 70$		
v	Data		p	Data		p	Data		p
Red	0.905	1	0.046	0.893	1	0.046	0.537	1	*0.001*
Blue	0.612	1	*0.001*	0.302	0	*0.003*	0.071	0	0.026
Avg.	0.767			0.578			0.317		
Manipulation (Part Two)									
	$k = 10$			$k = 40$			$k = 70$		
v	Data		p	Data		p	Data		p
Red	0.921	1	0.084	0.896	1	0.084	0.443	0	*0.000*
Blue	0.415	0	*0.002*	0.091	0	0.084	0.086	0	0.084
Avg.	0.772			0.650			0.328		

Tests on Freq. of Implementation (Human Researcher)

Model Prediction

v		$k_1 = 10$	$k_2 = 40$	$k_3 = 70$
Red	Manipulation	accept	accept	**reject**
	No-Manipulation	accept	accept	**accept**
Blue	Manipulation	**reject**	reject	reject
	No-Manipulation	**accept**	reject	reject

p-value for two-tailed matched-pair Signed Rank Tests (18 obs.)

	$k = 10$	$k = 40$	$k = 70$
Red vs. Blue (no-manipulation)	*0.003*	0.000	0.002
Red vs. Blue (Manipulation)	0.002	0.000	*0.002*
No-manipulation vs. Manipulation (Red)	0.979	0.968	*0.184*
No-manipulation vs. Manipulation (Blue)	*0.274*	*0.036*	0.547

Evaluators' Freq. of Implementation (Robot Researcher)

No-manipulation (Part One)									
	$k = 10$			$k = 40$			$k = 70$		
v	Data		p	Data		p	Data		p
Red	0.978	1	0.317	0.926	1	0.084	0.659	1	*0.002*
Blue	0.868	1	0.026	0.198	0	*0.005*	0.095	0	0.084
Average	0.922			0.578			0.361		
Manipulation (Part Two)									
	$k = 10$			$k = 40$			$k = 70$		
v	Data		p	Data		p	Data		p
Red	0.978	1	0.084	0.993	1	0.317	0.438	0	*0.002*
Blue	0.409	0	*0.005*	0.146	0	0.026	0.020	0	0.317
Average	0.839			0.800			0.322		

Tests on Freq. of Implementation (Robot Researcher)

Model Prediction

v		$k_1 = 10$	$k_2 = 40$	$k_3 = 70$
Red	Manipulation	accept	accept	**reject**
	No-manipulation	accept	accept	**accept**
Blue	Manipulation	**reject**	reject	reject
	No-manipulation	**accept**	reject	reject

p-value for two-tailed matched-pair Signed Rank Tests (18 obs.)

	$k = 10$	$k = 40$	$k = 70$
Red vs. Blue (No-manipulation)	0.105	0.000	0.002
Red vs. Blue (Manipulation)	0.001	0.000	*0.003*
No-manipulation vs. Manipulation (Red)	0.564	0.084	0.037
No-manipulation vs. Manipulation (Blue)	0.004	0.407	0.564

Summary of Evaluators' Behavior

Combining Finding 2 and 3, the experimental data is overall consistent with the theory predictions.

- The theory predictions are point and extreme predictions (0 or 1 predictions), so any noise /experimentation/confusion can be deviation from the theory.
- Comparative statics is more important to evaluate the theory than the point predictions.

Summary of Evaluators' Behavior Cont'd

p-value comparing Human and Robot Researcher treatments

	No-manipulation (Part One)		
	$k = 10$	$k = 40$	$k = 70$
Red	0.171	0.598	0.325
Blue	0.008	0.572	0.528

	Manipulation (Part Two)		
	$k = 10$	$k = 40$	$k = 70$
Red	0.865	0.258	0.732
Blue	0.631	0.432	0.324

Finding 4: *Overall, Evaluators' frequency of implementation is not significantly different between Human Researcher and Robot Researcher treatments.*

Welfare Comparison: Manipulation vs. No-manipulation

- Researcher's welfare:
  - When k=10, no difference (p=0.53): contrast to theory
  - When k=40, increased under Manipulation (p=0.03): consistent with theory
  - When k=70, increased under Manipulation (p=0.05): contrast to theory

Welfare Comparison: Manipulation vs. No-manipulation

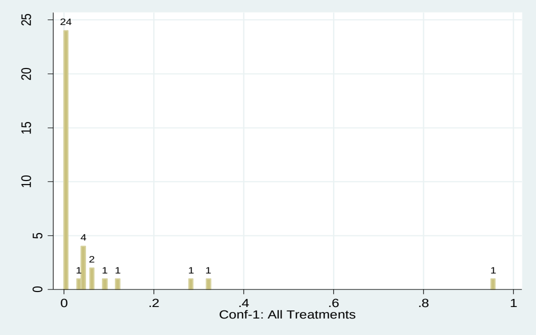
- Researcher's welfare:
  - When k=10, no difference (p=0.53): contrast to theory
  - When k=40, increased under Manipulation (p=0.03): consistent with theory
  - When k=70, increased under Manipulation (p=0.05): contrast to theory
- Evaluator's welfare:
  - When k=10, increased under Manipulation (p=0.005): consistent with theory
  - When k=40, increased under Manipulation (p=0.001): consistent with theory
  - When k=70, decreased under Manipulation (p=0.004): consistent with theory

Discussion: Explanations on deviation from the theory

- Strategy uncertainty and other-regarding preference are not the explanation
- Risk aversion alone cannot explain all the deviations from predictions
- Subjects may be confused
- Subjects may not use Bayesian updating on beliefs

Discussion: Explanations on deviation from the theory

- If Evaluator chooses not to implement when $k = 10$ or $k = 40$ given Red evidence, he must be confused.
- Using data in these two cells, we calculate a confusion index for each individual Evaluator.



Conclusion

- We test experimentally a game-theoretical model of persuasion bias in research conduction.
- In the model, Researcher and Evaluator have conflicts of interest.
- Researcher may manipulate sample selection.
- We design the experiment to focus on the behaviour and welfare of both parties when such manipulation is possible or not.
- We also compare treatments in which whether human subjects or robots play in the role of Researcher.

Conclusion Cont'd

- We find Researcher's behaviour is mostly consistent with theory, but there are significant deviations of Evaluator's behaviour from theory predictions.
- However, the comparative statics are still consistent with theory.
- No significant differences found between Human Researcher and Robot Researcher treatments.
- In the welfare analysis, we find Researcher is not worse off when manipulating, but Evaluator is harmed when k is large.

Conclusion Cont'd

For future research:

- A multiple-discipline approach may answer the questions more comprehensively
- Behavioral models which incorporate reputation concerns, researchers' social responsibility, positive externality of research outcomes may be considered

Procedure for Welfare Calculation

- Actual realizations of random events are different across treatments, and the actual frequencies are different from the expected probabilities assumed by theory.
- Therefore, it is difficult to conduct fair comparisons using the actual payoffs, which depend on the actual realizations of random events.
- We propose a procedure to calculate a welfare index that uses the expected probabilities but the actual choices of subjects, in order to remove the effect of different realizations of random events across treatments.

Procedure for Welfare Calculation Cont'd

- Each Researcher's welfare index depends on
  - session-level avg. of individual Evaluators' freq. of acceptance given v and k
  - Researcher's individual freq. of using Intuitive Strategy
  - ex-ante probability of random events

Procedure for Welfare Calculation Cont'd

- Each Researcher's welfare index depends on
  - session-level avg. of individual Evaluators' freq. of acceptance given v and k
  - Researcher's individual freq. of using Intuitive Strategy
  - ex-ante probability of random events
- Each Evaluator's welfare index depends on
  - session-level avg. of individual Researchers' freq. of using Intuitive Strategy
  - Evaluator's individual freq. of acceptance given v and k
  - ex-ante probability of random events