On the Fragility of Mediation

Alessandra Casella, Evan Friedman, and Manuel Perez Archila

November 2025

Introduction

- An experimental study on the effectiveness of mediation.
- Increasing interest in mediation: families, labor relations, legal disputes.
- Emergence of mediation algorithms.
- *Mediator*: a third party who wants to minimize conflict but has no independent resources, no superior information, and no enforcement power.

Main question:

Theory predicts that a benevolent mediator can strictly increase the probability of a peaceful resolution of conflict.

- A famous result in mechanism design (Myerson).
- A large literature, theoretical and empirical, debating when the result holds.
- Is this true in the lab?
- A good question for the lab: use a computerized optimal mediation mechanism.

The model: Hörner, Morelli, Squintani, 2015

- Two risk-neutral players compete for a resource of size 1.
- If they do not agree, the resource shrinks to $1/2 < \theta < 1$,
- and is divided according to the two players' types, H or L:

  $\theta/2$ to each if types are equal;

  θ to H and 0 to L otherwise.

- Ex ante efficiency corresponds to maximizing the probability of peaceful resolution.

- Players' types are private information and assigned independently.
- Each is H with known probability q, L with probability $1 - q$.

- Players attempt to reach agreement via a two-stage game:
  1. A communication stage. Cheap talk.
  2. A demand or allocation stage.
- Two procedures:
  1. Direct (unmediated) communication: DC
  2. Mediated communication: MC.

Unmediated communication (DC)

- Knowing one's own type t, each player sends to the other player a message $m \in \{s, h, l\}$.
- After messages are sent and received, each player expresses a demand $d \in \{1-\theta, 1/2, \theta, w\}$.
- If neither player chooses w and $d_1 + d_2 \leq 1$, each receives d_i.
- If either player chooses w, or if $d_1 + d_2 > 1$, the resource shrinks to θ and is divided according to the players' types.

Computer mediation (MC)

- The mediator M wants to maximize the probability of peace.
- M knows q but not the types' realizations.
- Each player sends M a confidential message $m \in \{s, h, l\}$.
- M recommends $r \in \{\{1/2, 1/2\}, \{\theta, 1-\theta\}, w\}$.
- If $r = w$ or if either player rejects r, the resource shrinks to θ and is divided according to the players' types.
- Otherwise, r is implemented.

The Myerson mediator

If M can commit to $r = w$ with positive probability, then:

Proposition HMS. *If* $(2\theta - 1) < q < (2\theta - 1)/\theta$, *mediation can achieve a strictly higher probability of peace than any equilibrium of the mediated communication game.*

Note:

- The mediator uses no transfers.
- The mediator has no superior information.
- The mediator has no enforcement power.

The confidentiality of the messages allows the mediator to "obfuscate" the opponent's type, induce H to accept $r = 1/2$, and keep L sincere while holding w lower.

Note: If H is not sure of the opponent's type,

- q cannot be so high that H always accepts $1/2$.
- q cannot be so low that H always prefers conflict to $1/2$.

Lying

In addition to higher peace, we also expect lower lying under MC than under DC.

Proposition 1. *Consider any equilibrium of the DC game in which $d = w$ is never played. If $\theta/2 > 1 - \theta$, then at least one type of player must be lying with strictly positive probability.*

Note:

- $d = w$ is weakly dominated by $d = \theta$.
- If all are truthful, L can strictly gain by sending message h.

But are the results likely to hold?

$\Longrightarrow$ An experiment.

The experiment

- $\theta = 0.7$.
- $q = 1/2$; $q = 1/3$. Fixed in each session.

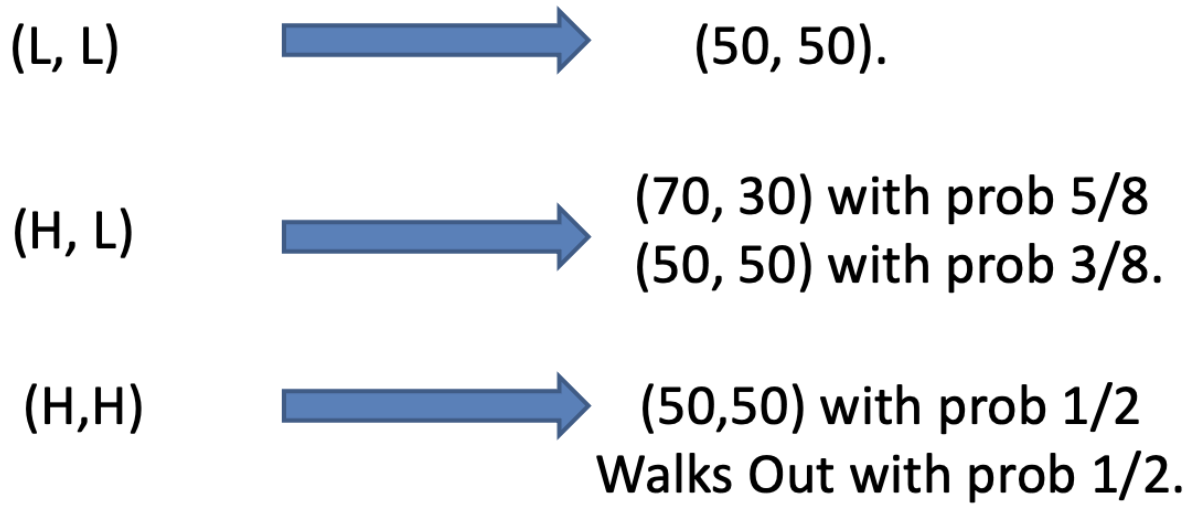
  $q = 1/2 \implies q > (2\theta - 1)$: Prop HMS applies.

  $q = 1/3 \implies q < (2\theta - 1)$: Prop HMS does not apply.

- Two treatments: DC and MC

HMS's optimal mediation program, on screen

The Computer Mediator's plan:



If the computer receives a Silent message from a player, it interprets it as either H or L with equal probability of 1/2 each.

Figure: *Computer mediator:* $q = 1/2$

The Computer Mediator's plan:

(L, L)	→	(50, 50).
(H, L)	→	(70, 30) with prob 3/4 Walks Out with prob 1/4.
(H,H)	→	Walks Out.

If the computer receives a Silent message from a player, it interprets it as an H with probability 1/3 and an L with probability 2/3.

Figure: *Computer mediator:* $q = 1/3$

- Two orders: DC, MC; or MC, DC.

  (Actually: DC, X, MC; or MC, X, DC)
- 20 rounds per treatment, with random matching.
- 3 sessions per parametrization per order: 12 sessions.
- 12 subjects per session; 144 subjects in total.

- Hypotheses:

  1. MC leads to more truthful reporting than DC.
  2. MC leads to more peace than DC.

L's are more sincere in MC

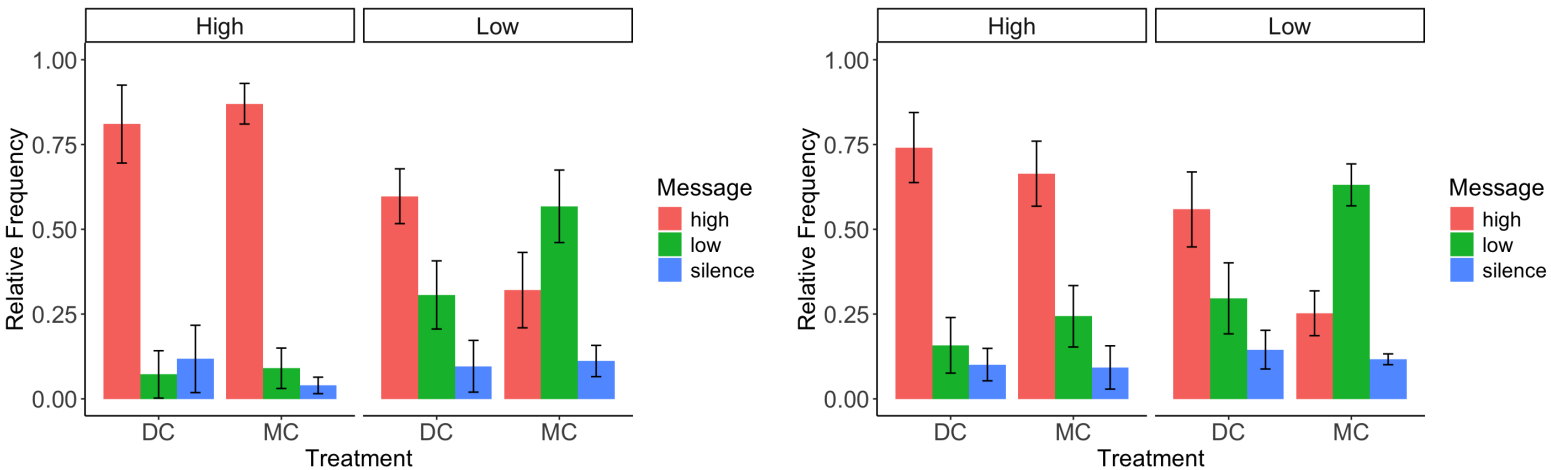


Figure: *Messages*. On the left: $q = 1/2$, on the right: $q = 1/3$. Standard errors are clustered at the session level.

But peace is not higher

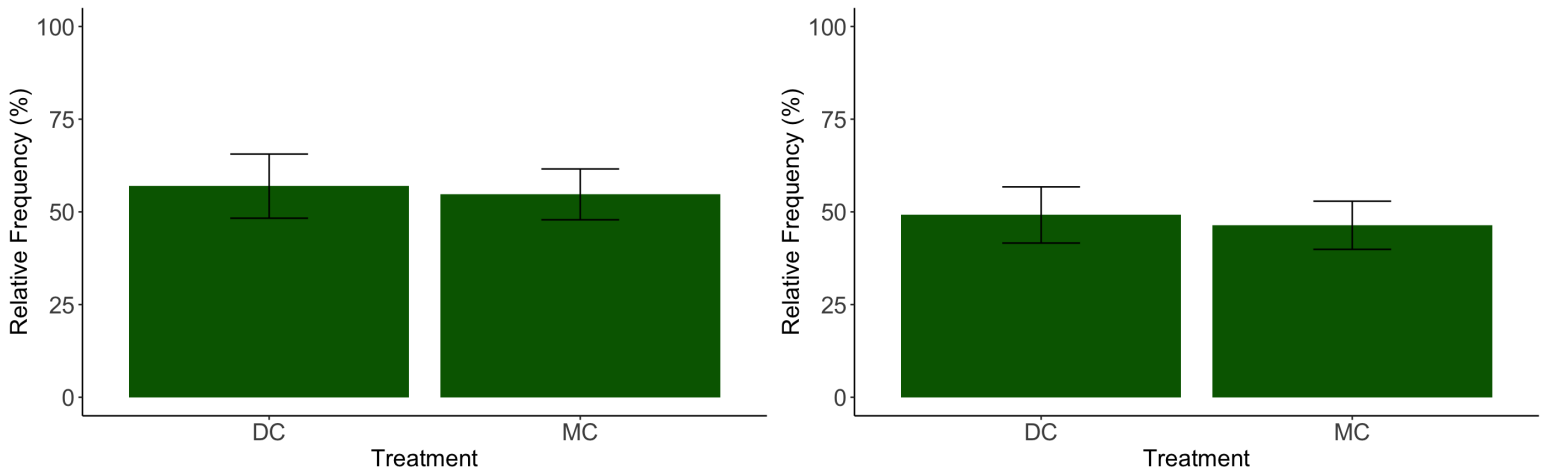


Figure: *Frequency of peace.* On the left: $q = 1/2$; on the right: $q = 1/3$. Standard errors are clustered at the session level.

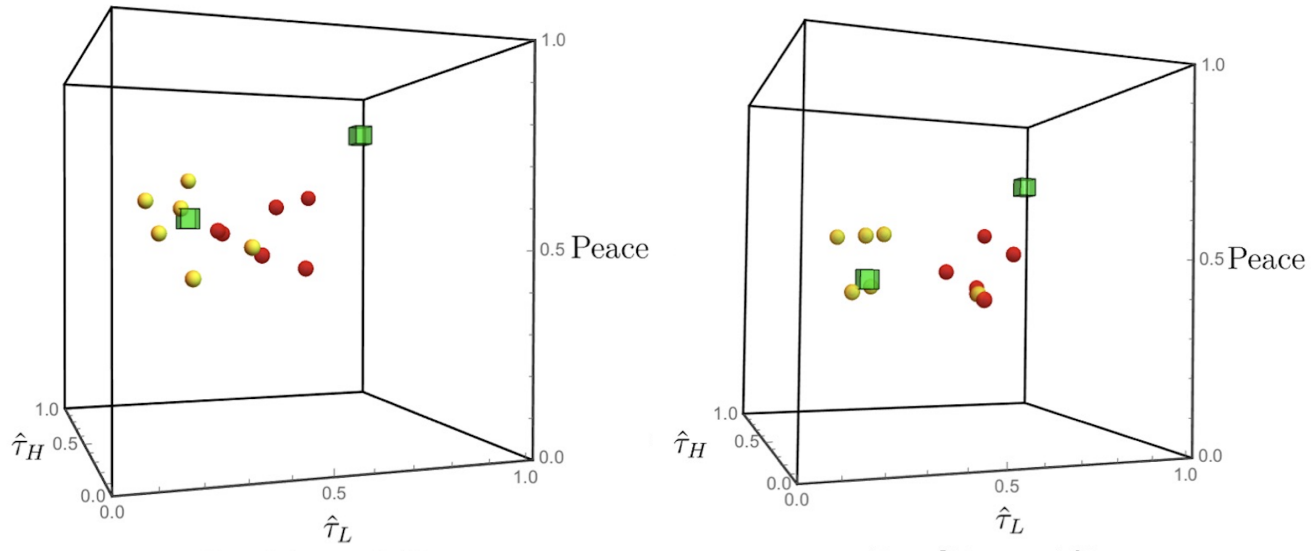


Figure: *Sincerity and peace by session.* On the left: $q = 1/2$; on the right: $q = 1/3$. Yellow: DC; Red: MC; Green: Theory.

The frequency of peace under MC falls short. Why?

Why Does Mediation Fall Short?

1. Multiple equilibria

- HMS characterize the "best" equilibrium.
- But keeping fixed the mediator's program, MC has many equilibria.

1. Multiple equilibria

- HMS characterize the "best" equilibrium.
- But keeping fixed the mediator's program, MC has many equilibria.
- We concentrate on equilibria in undominated strategies where, regardless of message:

  (i) all players accept 70;

  (ii) L players always accept 50;

  (iii) H players always reject 30.

- The equilibrium strategies to be determined are:

  (i) The acceptance strategies of *Hh* and *Hl* players offered 50, and of *Ll* players offered 30;

  (ii) The first stage message strategies for both types.

Selecting equilibria not grossly contradicted by the data: *LI* accepts 30; $\tau_H \geq \tau_L$.

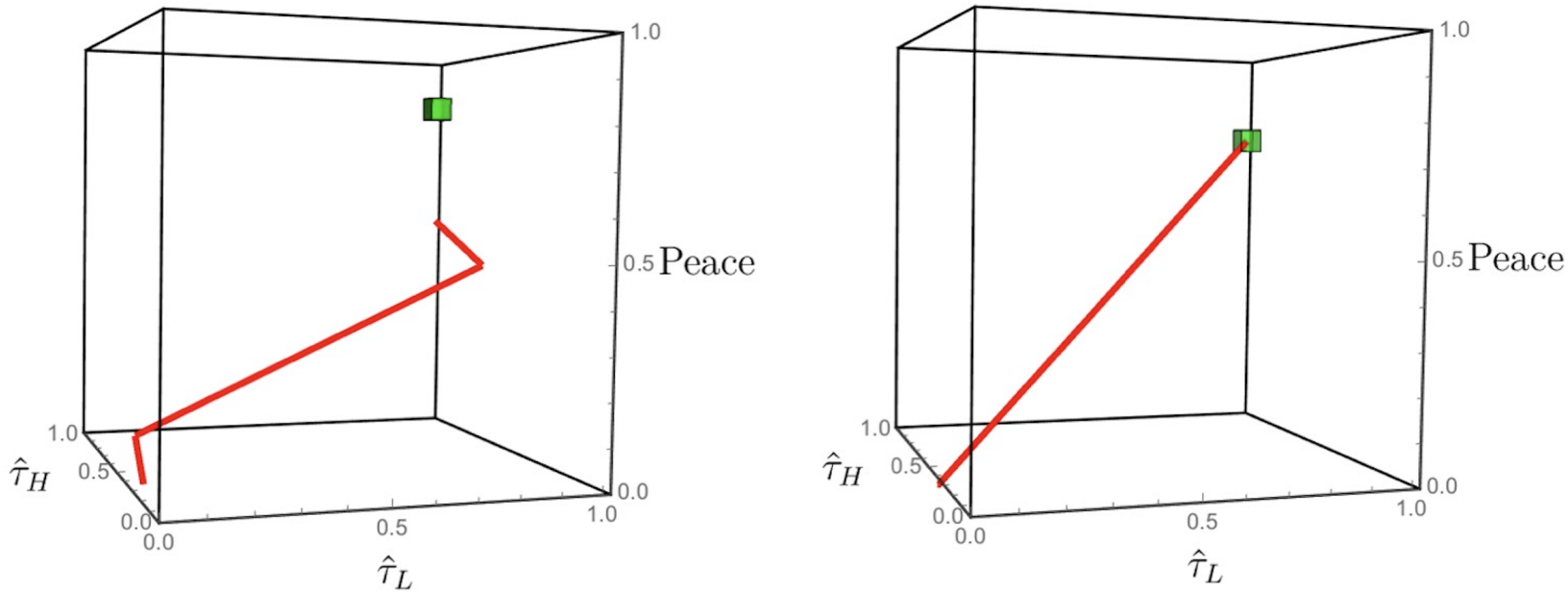


Figure: *Equilibria in undominated strategies.* On the left: $q = 1/2$; on the right: $q = 1/3$. The cube is the HMS equilibrium.

- Given the mediation program, equilibrium can support a large range of truthfulness and any peace between 0 and the HMS max.
- With $q = 1/2$, the locus of equilibria is discontinuous around the HMS equilibrium.

2. Fragility of the equilibrium with obfuscation

Call α_h the prob that an *Hh* player accepts 50.

Proposition 2. *Suppose* $(2\theta - 1)/\theta > q > (2\theta - 1) > 0$. *Then*:

(i) $\alpha_h = 1 \Longrightarrow \{\tau_H = 1, \tau_L = 1\}$.

(ii) *If either* $\tau_H < 1$ or $\tau_L < 1$, *then* $\alpha_h = 0$.

(iii) $\{\tau_H = 1, \tau_L = 1\} \nRightarrow \alpha_h = 1$.

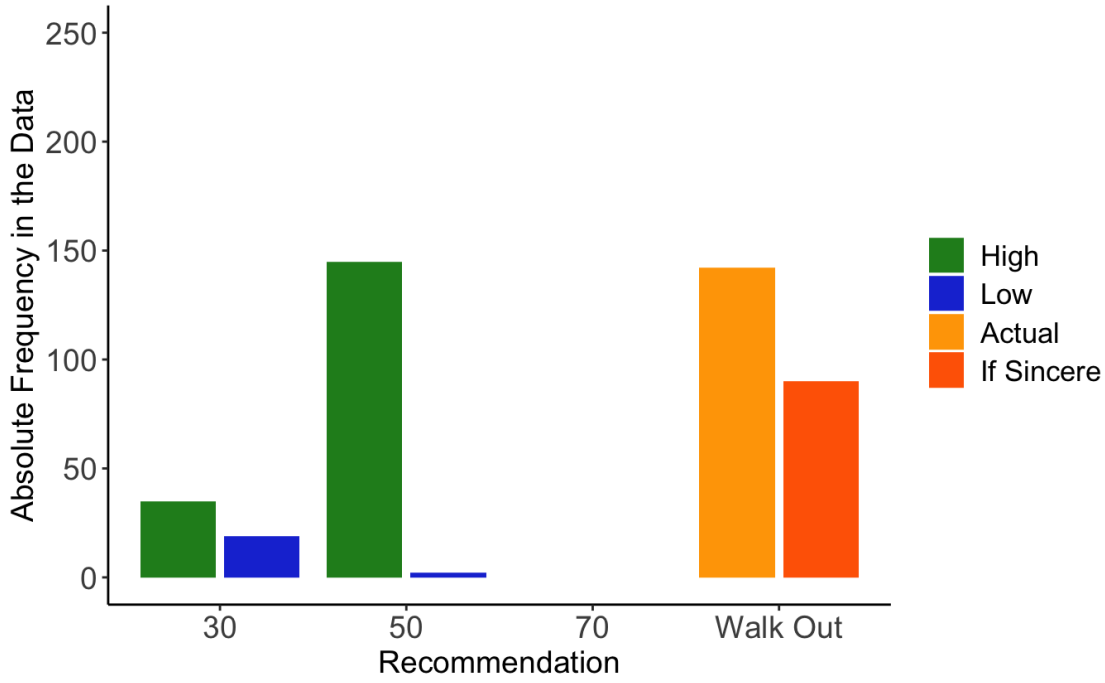


Figure: *Causes of conflict.* $q = 1/2$

- The discontinuity does not exist in the absence of obfuscation (the IR constraints are slack or bind trivially).
- Obfuscation is at the heart of mediation's superior effectiveness.
- But the equilibrium with obfuscation is fragile: it can only hold if both H and L types are fully sincere.
- With $q = 1/3$, optimal mediation has no obfuscation, and the locus of equilibria has no discontinuity at $\{\tau_L = 1, \tau_H = 1\}$.
- But it is steep and peace falls rapidly as sincerity decreases.

3. Deviations from equilibrium

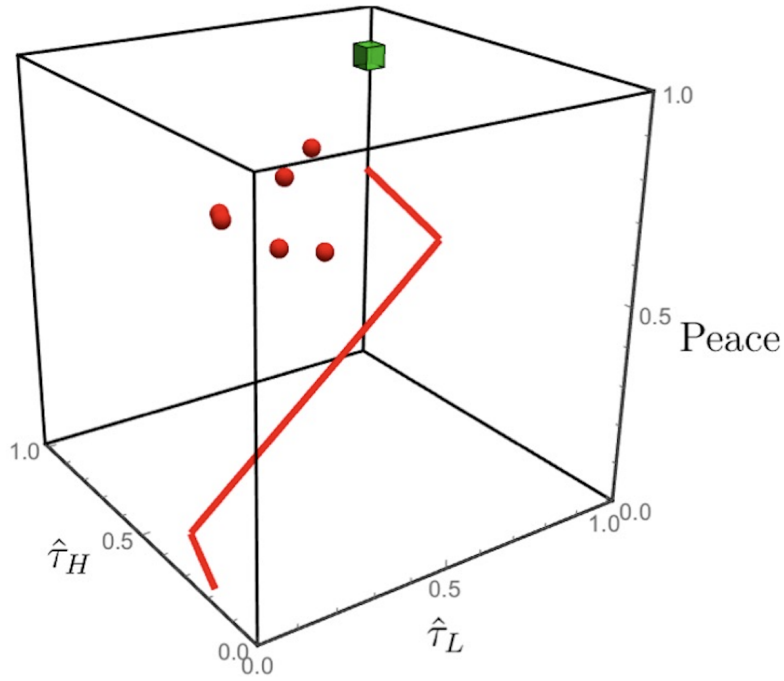


Figure: *Equilibria and experimental data.* $q = 1/2$

Under $q = 1/2$, deviations are due mostly to higher sincerity and compliance by H's.

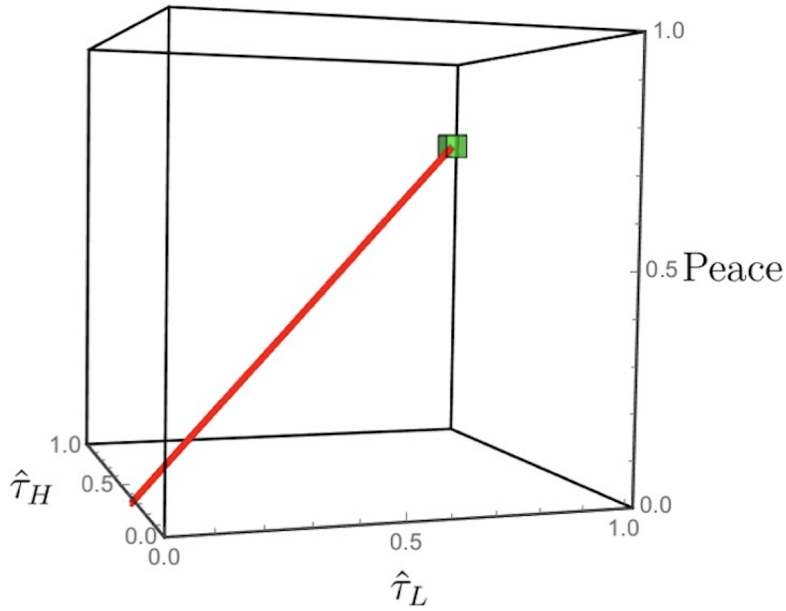


Figure: *Equilibria and experimental data.* $q = 1/3$.

Under $q = 1/3$, deviations are due mostly to imperfect sincerity by H's who then reject the offer.

3. Deviations are not costly

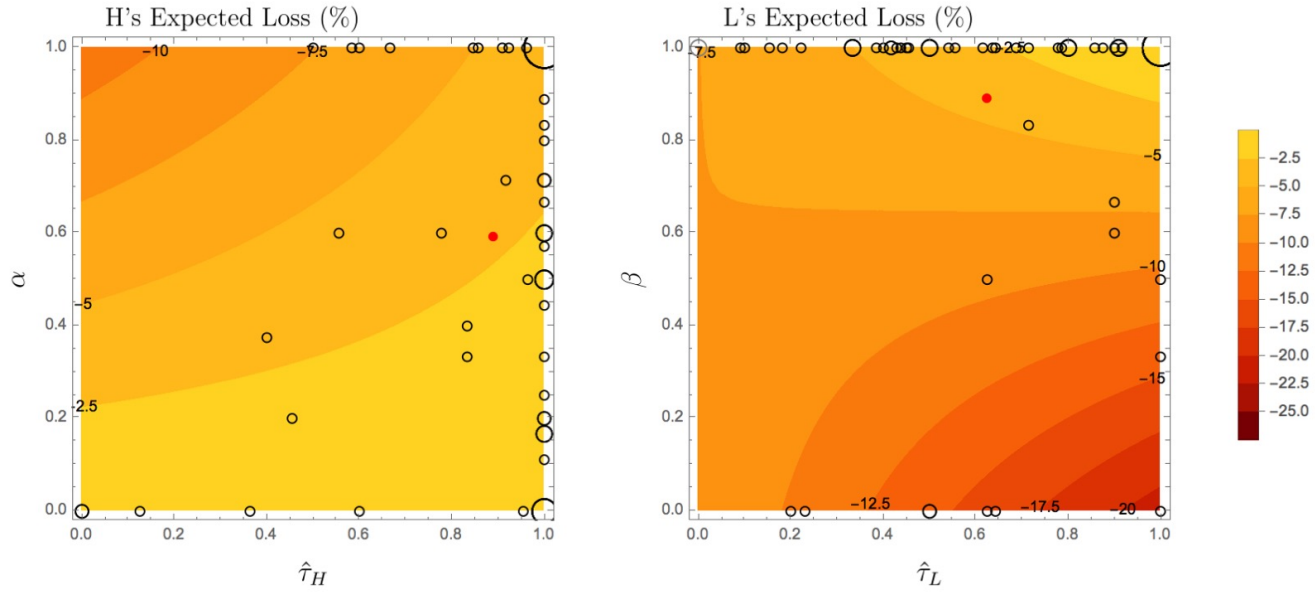


Figure: *Empirical losses given others' strategies.* $q = 1/2$

- $q = 1/2$. 93% of H's and 74% of L's lose less than 5%.
- Note: For H, the incentive constraints are not satisfied.

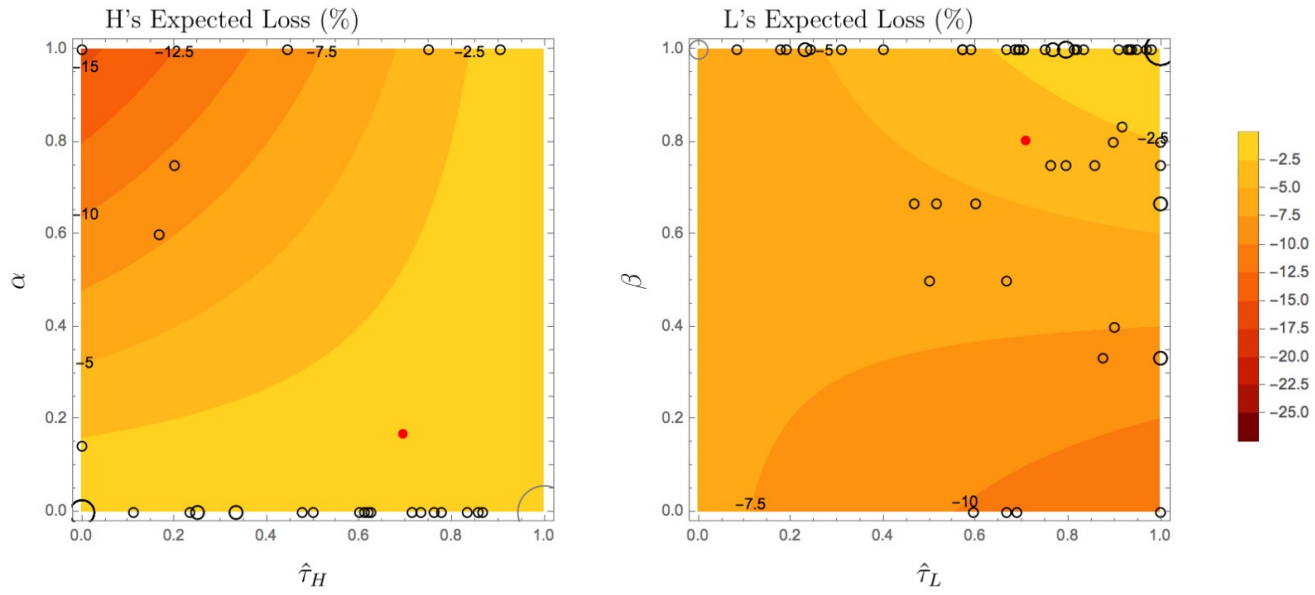
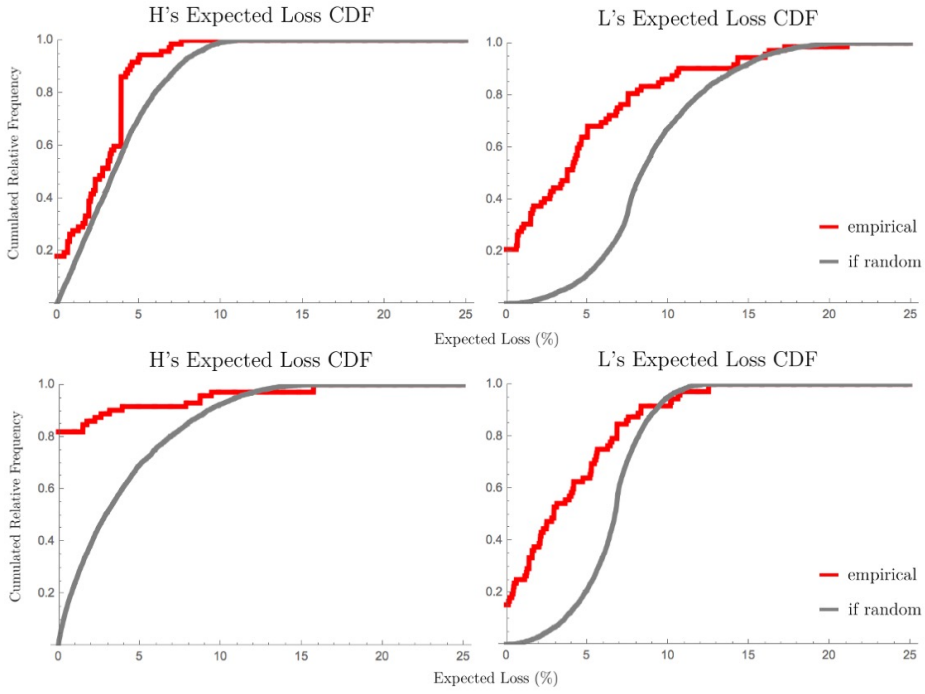


Figure: *Empirical losses given others' strategies.* $q = 1/3$

- $q = 1/3$. 92% of H's and 64% of L's lose less than 5%; 85% of L's lose less than 7.5%.
- Note: The incentive constraints are satisfied for both types.

- Subjects do not play erratically (KS tests).
- And indeed losses are lower.



Panel A: q = 1/2

Panel B: q = 1/3

Figure: CDFs of Losses Relative to Best Responding: Data vs Random Decision-Making

Preliminary conclusions

- In the theory and in the lab, fragility of the obfuscation equilibrium:
  1. Discontinuity around the HMS equilibrium if the mediation program includes obfuscation. Not otherwise.
  2. In the data, compliance and sincerity are best responses for H if the mediation program does *not* include obfuscation. Not otherwise.

- Yet both mediation programs fall short of their best theoretical promise:
  1. Multiple equilibria.
  2. Actions with small individual costs induce high conflict.
- The optimal mediation program is not optimal in the lab.

  $\Longrightarrow$ Design mechanisms with slack.

Adding slack to the mediation program

- In the optimal equilibrium, IC constraints hold weakly.
- We can make the equilibrium with sincerity and compliance strict.

For $q = \frac{1}{2}$, $d \in [0, \frac{3}{8})$ and $e \in [0, \frac{1}{2})$:

- (l, l) offered (50,50) w.p. 1
- (h, l) offered (70,30) w.p. $5/8 + d$; (50,50) w.p. $3/8 - d$
- (h, h) offered (50,50) w.p. $1/2 - e$; w w.p. $1/2 + e$.

- If $d = e = 0$, the mechanism coincides with HMS.
- If $\frac{4}{5}d < e < \frac{4}{3}d$, there is an equil with full sincerity and compliance, and strict IC's.
- In such an equilibrium, however, the discontinuity at full sincerity remains.

Robust mechanism in the lab

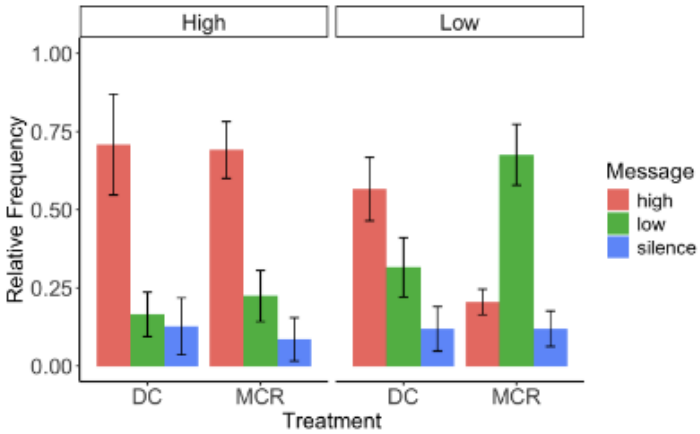
For $q = \frac{1}{2}$: $P = 0.825$; $(P(DC) = 0.59)$.

- (l, l) offered (50,50) w.p. 1
- (h, l) offered (70,30) w.p. 4/5; (50,50) w.p. 1/5
- (h, h) offered (50,50) w.p. 3/10; w w.p. 7/10.

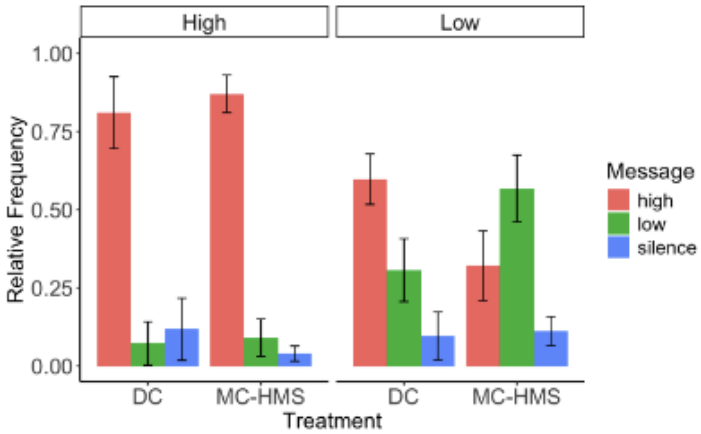
For $q = 1/3$: $P = 0.68$; $(P(DC) = 0.44)$.

- (l, l) offered (50,50) w.p. 1
- (h, l) offered (70,30) w.p. 1/2; w w.p. 1/2
- (h, h) offered (50,50) w.p. 1/8; w w.p. 7/8.

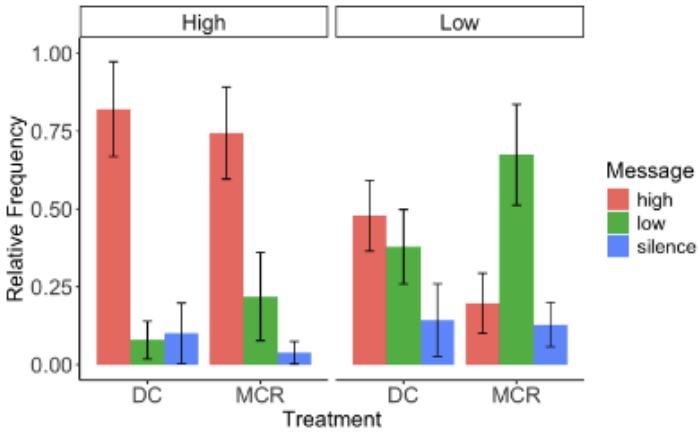
Sincerity under robust mechanism



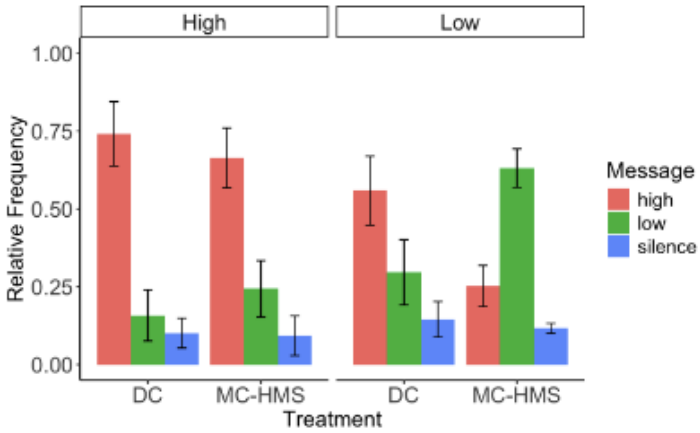
(a) DC and MCR; $q = 1/2$. Auxiliary sessions.



(b) DC and MC-HMS; $q = 1/2$. Original sessions.



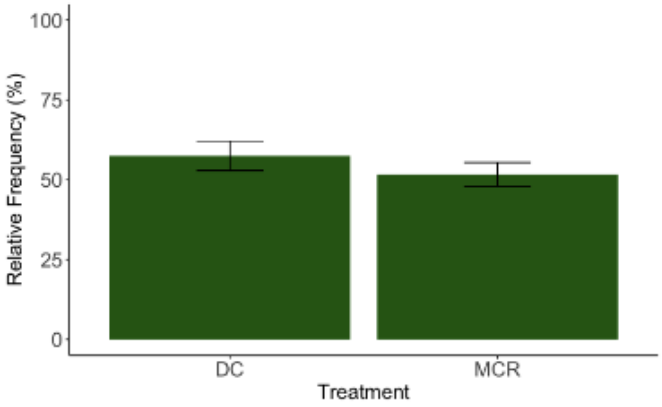
(c) DC and MCR; $q = 1/3$. Auxiliary sessions.



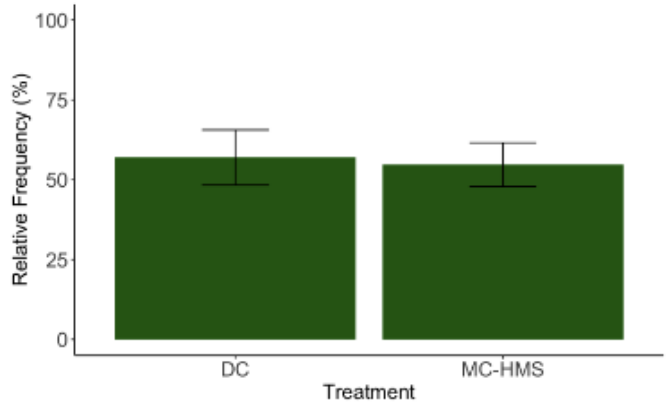
(d) DC and MC-HMS; $q = 1/3$. Original sessions.

Figure: *Sincerity. Robust MC and MC HMS*

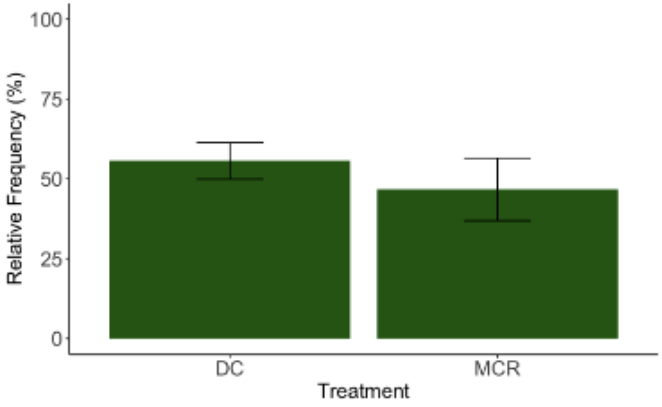
Peace under robust mechanism



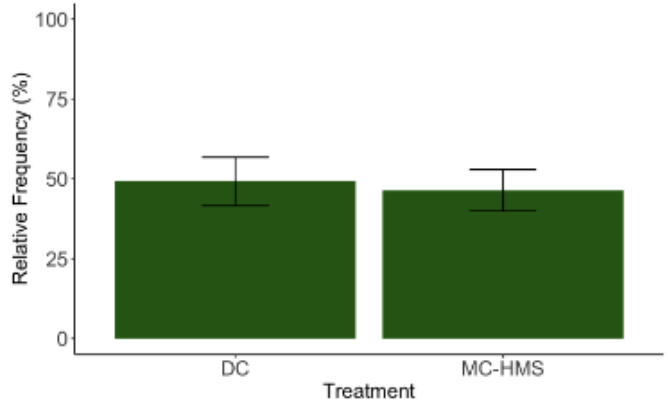
(a) DC and MCR; $q = 1/2$. Auxiliary sessions.



(b) DC and MC-HMS; $q = 1/2$. Original sessions.



(c) DC and MCR; $q = 1/3$. Auxiliary sessions.



(d) DC and MC-HMS; $q = 1/3$. Original sessions.

Figure: *Peace. Robust MC and MC HMS.*

Conclusions

- In the lab, the optimal mediation mechanism performs weakly.
- Multiple equilibria and fragilities to noise make the mechanism vulnerable.
- There is need for a robust mechanism. It must go beyond making IC constraints strict.