NBER WORKING PAPER SERIES

MEDIATING CONFLICT IN THE LAB

Alessandra Casella
Evan Friedman
Manuel Perez Archila

Working Paper 28137
http://www.nber.org/papers/w28137

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2020

We are extremely grateful to Massimo Morelli for his advice and financial support through ERC grant 694583. We thank participants at seminars at PSE, Essex, Columbia, Montréal, VIBES, the George Mason Law School, the 2020 Conference on Mechanism and Institution Design, and the 2020 SITE-PE meeting for their comments. In particular, we thank Marina Agranov, Marco Battaglini, Pedro Dal Bó, Sean Horan, Jacopo Perego, Erik Snowberg, and Leeat Yariv for their detailed reactions and suggestions. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Mediating Conflict in the Lab
Alessandra Casella, Evan Friedman, and Manuel Perez Archila
NBER Working Paper No. 28137
November 2020
JEL No. C78,C92,D74,D82,D86

ABSTRACT

Mechanism design teaches us that a mediator can strictly improve the chances of peace between two opponents even when the mediator has no independent resources, is less informed than the two parties, and has no enforcement power. We test the theory in a lab experiment where two subjects negotiate how to share a resource; in case of conflict, the subjects' privately known strength determines their payoffs. The subjects send cheap talk messages about their strength to one another (in the treatment with direct communication) or to the mediator (in the mediation treatment), before making their demands or receiving the mediator's recommendations. We find that, in line with the theory, messages are significantly more sincere when sent to the mediator. However, contrary to the theory, peaceful resolution is not more frequent, even when the mediator is a computer implementing the optimal mediation program. While the theoretical result refers to the best (i.e. most peaceful) equilibrium under mediation, multiple equilibria exist, and the best equilibrium is particularly vulnerable to small deviations from full truthfulness. Subjects are not erratic and their deviations induce only small losses in payoffs, and yet they translate into significant increases in conflict.

Alessandra Casella
Department of Economics
Columbia University
420 West 118 Street
New York, NY 10027
and NBER
ac186@columbia.edu

Manuel Perez Archila
International Monetary Fund
HQ1 6-318.1
700 19th Street, NW
Washington, DC 20431
mperez-archila@imf.org

Evan Friedman
Department of Economics
University of Essex
Wivenhoe Park
Colchester CO4 3SQ
United Kingdom
ekf2119@columbia.edu

A data appendix is available at http://www.nber.org/data-appendix/w28137

Mediating Conflict in the Lab*

Alessandra Casella† Evan Friedman‡ Manuel Perez Archila§

November 6, 2020

Abstract

Mechanism design teaches us that a mediator can strictly improve the chances of peace between two opponents even when the mediator has no independent resources, is less informed than the two parties, and has no enforcement power. We test the theory in a lab experiment where two subjects negotiate how to share a resource; in case of conflict, the subjects' privately known strength determines their payoffs. The subjects send cheap talk messages about their strength to one another (in the treatment with direct communication) or to the mediator (in the mediation treatment), before making their demands or receiving the mediator's recommendations. We find that, in line with the theory, messages are significantly more sincere when sent to the mediator. However, contrary to the theory, peaceful resolution is not more frequent, even when the mediator is a computer implementing the optimal mediation program. While the theoretical result refers to the best (i.e. most peaceful) equilibrium under mediation, multiple equilibria exist, and the best equilibrium is particularly vulnerable to small deviations from full truthfulness. Subjects are not erratic and their deviations induce only small losses in payoffs, and yet they translate into significant increases in conflict.

1 Introduction

Consider two parties fighting over a contested resource. The two parties can try to find an agreement directly or can resort to a *mediator*, an impartial third party who can hear their claims and make a recommendation but has neither independent resources nor power of enforcement. Can the mediator help to resolve the conflict?

From family disputes to labor relations to international conflict, mediation continues to attract sustained and indeed increasing attention, from psychologists, lawyers and judges, international or-

*We thank participants at seminars at PSE, Essex, Columbia, Montréal, VIBES, the George Mason Law School, the 2020 Conference on Mechanism and Institution Design, and the 2020 SITE-PE meeting for their comments. In particular, we thank Marina Agranov, Marco Battaglini, Pedro Dal Bó, Sean Horan, Jacopo Perego, Erik Snowberg, and Leeat Yariv for their detailed reactions and suggestions. We are extremely grateful to Massimo Morelli for his advice and financial support through ERC grant 694583.

†Columbia University, NBER and CEPR, ac186@columbia.edu.

‡University of Essex, ef20396@essex.ac.uk

§Columbia University, mfp2122@columbia.edu

ganizations, and corporate agents specializing in its craft.[1] Concentrating attention to international conflict, widely quoted surveys conclude that "half of all interstate wars and one third of all international crises since World War I have involved mediation" (Beardsley, 2011, p.3); Wilkenfeld et al. (2003) find that the recourse to mediation in international crises amounted to almost two thirds in 1990-96 (p. 286).

If the mediator has access to additional resources or can issue credible threats, or if the mediator has superior information about the parties' preferences and strengths, then we can predict that the mediator's involvement will matter. But if that is not the case? The surprising answer, proposed by Myerson in a particularly elegant application of mechanism design, is that a mediator adjudicating a dispute between self-interested and rational parties can still increase the chances of peace, in the absence of transfers, information, or power (Myerson, 1991, ch.6). The essence is the confidentiality of the communication between the contestants and the mediator. It is possible for the mediator to induce the parties to reveal their actual strength, and yet issue recommendations that leave them uncertain about the strength of their opponent and thus willing to accept the recommendation rather than engage in open conflict. The final result is a higher frequency of peaceful resolutions than the two sides can obtain by communicating directly, a higher frequency fully engineered by the mediator's subtle modulation of incentives in the recommendations issued. As Myerson phrases it, the key is the *obfuscation* the mediator can employ–the possibility of not fully revealing the opponent's message–obfuscation that the parties cannot achieve by direct communication.

The optimal mediation mechanism is of interest both as a seminal theoretical result, and for what it may teach us for practical applications in actual disputes. This study presents an experimental test of Myerson's mediator. More broadly, it adds to the literature on the robustness of optimal mechanisms to multiple equilibria and noisy behavior.

Hörner, Morelli and Squintani (2015) exploited Myerson's insight in a model that confirms the superior role of mediation and that provides the game played in our experiment. Two players negotiate how to share a resource. In case of conflict, the players' privately known strength determines their payoffs. The players send cheap talk messages about their strength either to one another, in the direct communication treatment, or to the mediator, in the mediation treatment, before making their demands or receiving the mediator's recommended allocation. The mediator is a neutral third party

[1] See for example a recent *Financial Times* article: "Industry of peacemakers capitalizes on global conflict" (Oct. 22, 2019).

whose goal is to maximize the probability of a peaceful resolution, but who has no information about the players' levels of strength (beyond a common prior) and no enforcement power. Upon seeing the players' messages, the mediator can issue a recommendation or refuse to mediate. Peace prevails if the mediator accepts to mediate and if the recommendation is individually accepted by both players. In line with Myerson's lesson, Hörner et al. show that if the mediator can commit to refuse mediation under some conditions, then the ex ante probability of peace can be strictly higher than what the two players can achieve without a third party.[2]

The streamlined game lends itself well to experimental testing. We take it to the lab, with minimal departures from the theoretical model. The core of our experiment is the comparison between a treatment with direct communication and a mediation treatment where the mediator is a computer, known to enact the theoretical optimal mediation mechanism identified by Hörner et al..[3] We find that mediation does indeed increase sincerity, something that theory predicts in our setting: in particular, the possibility to send confidential messages is associated with higher willingness to admit weakness. However, mediation does not increase the frequency of peace.

There are several reasons why mediation does not fulfill its promise in the lab. First, keeping constant the mediation program, there is a multiplicity of equilibria which vary greatly both in the probability of truthful messages and in the frequency of peace. This is a well-known problem in mechanism design,[4] and not too surprisingly the lab makes it salient. Second, and more closely tied to the specific mechanism studied here, when the equilibrium with highest peace requires obfuscation, the locus of equilibria is discontinuous in the neighborhood of full truthfulness. With any deviation from full truthfulness, no matter how small, non-compliance with some of the mediator's recommendations becomes optimal, triggering a discrete upward jump in the probability of conflict. The fragility of the equilibrium with obfuscation matters because the theory teaches us that it is exactly the possibility of obfuscation that makes mediation superior to direct negotiations. In the data, the extent of non-compliance is less than the equilibrium analysis predicts, but is large enough to make the frequency of peace under mediation and under direct communication fully comparable.

Finally, the fragility of obfuscation is not the only source of increased conflict in the lab. In the absence of obfuscation too, the frequency of peace remains lower than the theory predicts under

[2] Hörner et al. reach a second strong result: in their model a mediator with no enforcement power is just as effective in avoiding conflict as an arbitrator who can force the two factions to accept the arbitration resolutions (but cannot force them to enter into the arbitration process).

[3] The optimal mechanism–the precise mapping from the pair of messages received and the mediator's recommendations–is publicly announced and known to the experimental subjects.

[4] See, for example, Palfrey (1990).

optimal mediation. Again, part of the reason is the multiplicity of equilibria, but another part is deviations from equilibrium behavior. And yet, even though the game is strategically complex, subjects' decisions are not erratic: for both messages and acceptance choices we strongly reject the hypothesis that actions are random. In the lab, observed actions are not always best responses to the choices of others, but the deviations consistently come with small payoff effects. This then is the third reason for our results. Optimal mediation is fragile not only because of multiple equilibria, not only because of the instability of the obfuscation equilibrium to small deviations from full truthfulness, but more broadly because a large set of strategies induces both small individual losses and high conflict.

Our study can be read under two alternative perspectives. It is primarily a contribution to the literature on theoretical mechanisms for bargaining and dispute resolution. The comparison of mediation to direct communication is the subject of a rich stream of theoretical works. Its authors find that the answer is sensitive to the details of the game: how long the direct communication can last (Forges, 1986; Aumann and Hart, 2003); whether it is only verbal or can take other forms (Forges, 1990; Krishna, 2007); whether the asymmetry of information is one or two-sided (Goltsman et al., 2009); whether, after the communication stage, the bargaining is one-shot or dynamic (Fanning, 2019, and the papers cited there). In a model very similar to that of Hörner et al., Fey and Ramsay (2010) find that mediation cannot improve over direct communication if the asymmetry of information concerns a private value–the idiosyncratic cost of conflict–as opposed to an interdependent value as in Hörner et al.–the strength of each party, and hence the probability of victory in case of conflict. On the experimental side, however, the contributions are very few. To our knowledge, the only other experiment testing a theoretical mediation mechanism is Blume et al. (2019), comparing mediation and direct communication in a sender-receiver game with one-sided asymmetric information. By design the mediation mechanism is sub-optimal but, in line with the optimal mechanism, garbles messages and allows for truthful revelation, while the only equilibrium with unmediated communication is an inefficient pooling equilibrium. In the lab, deviations from equilibrium prevent the full gains from mediation, but mediation does increase truthful revelation and leads to moderate payoff improvements.

Beyond the specific focus on mediation, our work tests the ability of experimental participants to use sophisticated strategies to convey and extract information in the lab. It recalls recent experimental studies on Bayesian persuasion (Frechette et al., 2019; Nguyen, 2017; Au and Li, 2018; Aristidou et al., 2019). However, these studies have repeatedly found that the information design problem is particularly challenging for inexperienced subjects in the lab. The mediation game is much more

intuitive, and commitment to the optimal mediation mechanism is built into the computer mediator's actions.

Our study is closer to the tradition of experiments in mechanism design. Where mechanism design has been particularly influential (in matching mechanisms, for example, or spectrum auctions), the theory has been complemented by experimental studies that have tested and fine-tuned the final format.[5] In this perspective, some of our conclusions echo other studies. The fragility of mediation caused by the multiplicity of equilibria recalls, for example, the results of Cason et al. (2006), where such multiplicity is found to hamper the implementation of a desirable social choice. The lack of robustness to small noise in behavior is the focus of Aghion et al. (2018), confronting subgame perfect implementation with behavioral biases in the lab. On the other hand, the specific vulnerability of the equilibrium with obfuscation does not appear to have a precedent in the experimental literature. Besides the theoretical importance of obfuscation, the finding may matter for applications. For example, Meirowitz et al. (2019), again working with the Hörner et al. model, single out exactly the mediation mechanism with obfuscation as the one dispute resolution institution for international conflicts that would not lead to increased militarization and eventually increased conflict. Our results invite some caution.

Under this second, applied perspective, one immediate question is whether the stripped down model can be instructive in practical instances of mediation. In particular, the theory depends on the mediator's willingness to commit to abandon mediation. Is this a reasonable assumption? In the case of professional mediators, with long term reputations to preserve, the answer seems positive. For example, in a highly cited article targeted to law practitioners, Brown and Ayres (1994) discuss in detail concrete means through which such commitment can be achieved. With respect to international conflict, Hörner et al. defend the empirical relevance of the assumption in their online appendix. Other modeling choices can be, and are, debated as well. The literature on mediation and international disputes often endows the mediator with independent information, as opposed to having to elicit information from the parties; the question then concerns whether such information can be conveyed accurately and believed, when the mediator is more or less biased towards one of the parties or towards peace (Kydd, 2003 and 2006, Rauchhaus, 2006; Smith and Stam, 2003). We do not take a position on this debate, besides stressing the usefulness of experiments in evaluating the theoretical results.

[5] For FCC auctions, see, for example, Banks et al., 2003, and Brunner et al., 2010. For matching mechanisms, see, among many others, Chen and Somnez, 2006; Roth, 2016. For VCG mechanisms for public good provision, see for example Attiyeh et al., 2000; Chen and Plott, 1996, and Chen, 2008.

From an experimental perspective, political scientists' works on mediation follow a methodology quite different from ours, less tightly tied to theory and closer to historical events: experiments simulate historical world crises and observe the impact of a mediator, trained to follow different protocols (see for example Wilkenfeld et al., 2003).

The paper proceeds as follows. The next section describes the model and its main theoretical properties, comparing optimal mediation, direct communication, and mediation in the absence of commitment power; Section 3 describes the experimental design; Section 4 reports the results; Section 5 discusses possible reasons why the optimal mediation mechanism is not more successful than direct communication in averting conflict in the lab. Section 6 reports equilibria for the direct communication treatment, as well as for an experimental treatment that implements mediation without commitment, a "human mediator" treatment, where the mediator was played by a third experimental participant. Finally, Section 7 concludes. Additional material, in particular but not exclusively the derivation of the theoretical equilibria, is collected in two Appendices.

2 The Model

The mediation game we took to the lab follows closely the model in Hörner, Morelli and Squintani (2015) (HMS), with a few modifications that streamline the experimental design. Two risk-neutral players, 1 and 2, compete for a resource of size 1. Each player is of type $T \in \{H, L\}$. Types are drawn independently for the two players and are private information, but it is commonly known that each player is of type H with probability q, and of type L with probability $1-q$. If 1 and 2 agree on sharing the resource peacefully, each receives the agreed share. If not, they go to war, the resource shrinks to $\theta < 1$ and is divided according to the two players' types: if the two players' types are equal, each receives $\theta/2$; if one player is H and the other is L, H receives the full amount θ and L receives 0. From an efficiency standpoint, distribution is irrelevant: maximizing ex ante efficiency corresponds to maximizing the probability of peaceful resolution.

An equal split $(1/2, 1/2)$ is always preferable to conflict for an L type; in the absence of other information, $(1/2, 1/2)$ is also acceptable to an H type if $1/2 \geq (1-q)\theta + q\theta/2$. To highlight the role of information, HMS (and we) assume $1/2 < (1-q)\theta + q\theta/2$, or $q < (2\theta - 1)/\theta$. In addition, we constrain $\theta/2 > 1 - \theta$ to ensure that the H type prefers to fight rather than accepting the smaller share when facing another H type.

The core of the analysis is the procedure through which the two players can reach an agreement. We consider two such procedures: unmediated (or direct) communication and mediation.[6] In both cases, the players take actions in two consecutive stages: a message stage and an allocation stage.

Under unmediated communication, after learning one's own type, at the message stage each player sends to the other player a cheap talk message $m(T)$. The message can be blank, or report a type as the player's own, but the report need not be truthful. Using lower case letters to indicate reported types, and s for the option to remain silent, $m \in \{s, h, l\}$. The two players send messages simultaneously. After messages are sent and received, the game moves to the allocation stage. At this stage, the two players, again moving simultaneously, express a demand $d(m, m', T)$, where m' stands for the opponent's message. Demand may consist of the refusal to negotiate, or indicate the demanded share of the resource. We constrain d to take one of four values: $d \in \{1-\theta, 1/2, \theta, w\}$, where w stands for "walking out", as we phrase it in the lab. If neither player chooses w and the two demands are compatible ($d_1 + d_2 \leq 1$), then each player receives what the player demanded, and peace prevails. If either player chooses w, or if $d_1 + d_2 > 1$, then no agreement is reached and war follows: the resource shrinks to θ and is divided according to the players' types.[7]

Under mediation, a third party enters the game, the mediator, whose objective is to maximize the probability of peace. The mediator shares the common prior q but has no information on the realizations of the players' types and has no enforcement power. At the message stage, each player sends the mediator a confidential message, where, as before, $m \in \{s, h, l\}$. On the basis of the messages received, the mediator recommends a division of the resource between the two players, or alternatively refuses to mediate. Denoting by r the mediator's recommendation, and indicating the share recommended to player 1 and then to player 2, in order, we constrain $r(m, m')$ to one of the following values $r \in \{(1-\theta, \theta), (1/2, 1/2), (\theta, 1-\theta), w\}$ where as before w stands for "walking out", or the mediator's refusal to mediate. If the mediator has made a recommendation, then, at the allocation stage, each player has the option of accepting the recommendation or rejecting it. The recommendation is implemented if both players accept it. If the mediator has refused to mediate, or if either player rejects the recommendation, then war follows, the resource shrinks to θ and is divided according to the players' types.

The mediator's ability to commit to war by refusing to mediate is essential to inducing players to

[6] HMS also consider arbitration, i.e. mediation with enforcement power.

[7] If $d_1 + d_2 < 1$, a third agent acquires what is left of the resource. In the lab, it is the experimenter by default.

be truthful in their messages. It is also key to the following result:

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 unmediated communication game.*

By the revelation principle, the optimal mediation program must result in a weakly higher frequency of peace than unmediated communication. But HMS' result is stronger: for parameters in the specified range, mediation can achieve a *strictly* higher probability.[8]

Under the optimal mediation program, there exists an equilibrium where all messages to the mediator reveal the player's type sincerely, and all mediator's recommendations are always accepted by the players. The two binding constraints are L's incentive compatibility constraint (L's incentive to be truthful), and H's ex post participation constraint (H's acceptance of the mediator's recommendation). The optimal mediation program has two crucial ingredients. First, the mediator refuses to mediate with positive probability following an h message (thus keeping L sincere)–the mediator is able to *commit*. Second, if $q > (2\theta - 1)$, the mediator's optimal recommendation does not reveal the opponent's type (thus limiting H's recourse to war when matched with an L)–although all messages are sincere, the opponent's type is *obfuscated*.

The mediator's ability to obfuscate explains the superiority of the optimal mediation equilibrium relative to what can be achieved under unmediated communication, where the direct messages prevent conditioning strategies on true types and yet having players uncertain about their opponent.[9] Effective obfuscation however requires a sufficiently high frequency of H types, high enough that an H player, uncertain about the type of the opponent, is still willing to accept an equal share of the resource. The equilibrium with obfuscation cannot be sustained if $q < (2\theta - 1)$. And in the absence of obfuscation, there exists an equilibrium of the direct communication game that replicates what mediation can accomplish.

As HMS show, the optimal equilibrium of the direct communication game is a correlated equilibrium that exploits a publicly observed randomization device. In our experimental design, there is no explicit public correlation device. In its absence, achieving the optimal equilibrium through the

[8] Our setting differs slightly from HMS' original model because we allow for silence and constrain both demands and the mediator's recommendations to lie in a restricted set. However, Proposition HMS continues to hold because the HMS equilibria remain optimal in our setting. First, under mediation, the option of remaining silent can be ignored (again by the revelation principle). Second, under both mediation and direct communication, the optimal HMS equilibria include only demands and recommendations in our restricted set of options. If the equilibria are optimal in the absence of restrictions, they must be optimal over the smaller set of programs that satisfy the restrictions.

[9] As noted in the Introduction, the superiority of mediation also depends on the interdependent nature of the variable subject to private information (see Fey and Ramsay, 2010).

randomization of individual messages is very difficult in principle,[10] and, we believe, impossible in practice. As a result, any equilibrium of the direct communication game played in the lab will be weakly suboptimal, and the highest achievable frequency of peace must be weakly lower. For emphasis, we state this observation as a separate Remark:

Remark: *In the equilibria of the unmediated communication game played in the lab, the probability of peace must be weakly lower than in the optimal equilibrium. Hence it must be lower than in the optimal mediation equilibrium, and strictly lower if* $(2\theta - 1) < q < (2\theta - 1)/\theta$.

The public correlation device exploited in the best equilibrium of the direct communication game induces war with a positive probability in response to specific pairs of messages. The possibility of war plays a disciplining role in equilibrium that mimics the commitment demanded from the mediator. In the lab, absent the public correlation device, the participants' option to walk-out ($d = w$) could in principle introduce commitment to war and induce sincerity. For example, there always exists a truthful equilibrium where all messages are sincere, all L types demand $1/2$, and all H types walk-out. In fact, as the next proposition shows, for the range of θ values that are relevant for the experiment ($\theta/2 > 1 - \theta$), full revelation in the unmediated communication game can *only* occur in equilibria where the option to walk-out is chosen with positive probability.

We focus, in the proposition below as in the rest of our analysis, on equilibria that are symmetric for players of given type. Then:

Proposition 1. *Suppose* $\theta/2 > 1 - \theta$*, and consider any equilibrium of the unmediated communication game in which* $d = w$ *is never played. Then at least one type of player must be lying with strictly positive probability.*

Proof. Suppose to the contrary that a fully revealing equilibrium exists where $d = w$ is never played. Consider the players' demand strategies, conditional on their type and their opponent's (fully revealed) type. Consider first a player of type T facing an opponent of the same type. With $\theta/2 > 1 - \theta$, war against an opponent of the same type yields more than $(1 - \theta)$; hence demanding $1 - \theta$ is strictly dominated.[11] Thus in any symmetric equilibrium with full revelation, in a match between two players of equal type, either both demand $1/2$, or both demand θ, or both mix between

[10] Mixed strategy profiles cannot typically result in correlated randomness (Forges (1986)). HMS make the same observation and document that, in the UC game, the optimal equilibria with mixed strategies at the message stage lead to a probability of war that under some parameters must be strictly higher than in the optimal correlated equilibrium with full truthfulness.

[11] The strategy is strictly dominated under the restriction that the opponent never plays $d = w$ (otherwise, the player's own demand could be irrelevant). The strategy is always weakly dominated.

1/2 and θ. Now consider a match between an H and an L. In such a match, the H player can always guarantee itself θ by asking for it, and the pair of demands $(\theta, 1-\theta)$ is the unique pair of mutual best responses. Consider then an L who reveals her type truthfully. If matched with an L, the highest possible realized share is 1/2; if matched with an H it is $(1-\theta)$. But then an L type has an incentive to deviate: declare type h, be believed, and best respond to the opponent's strategies. The L type masquerading as an H can demand and obtain θ against an L opponent, and at least $(1-\theta)$ against an H opponent. The deviation is strictly profitable. Hence a fully revealing equilibrium cannot exist. □

In the absence of repetition or commitment, $d = w$ is always weakly dominated by $d = \theta$, for any type and for any message sent and received. In the lab, random rematching induces lack of repetition, and in what follows we concentrate on symmetric equilibria in weakly undominated strategies, where $d = w$ is never chosen. We refer to them in short as "equilibria".[12]

In the experiment, optimal mediation is implemented by a computer algorithm, but we also investigated a more exploratory question: whether untrained experimental subjects could be effective mediators. The experimental design has anonymity and random rematching across rounds, with the result that mediators cannot build reputation and lack incentives to walk out of mediation. The absence of commitment power hinders the effectiveness of mediation. The following proposition, proved in the Appendix (Section 8.1), makes the case in the present model.

Proposition 2. *Assume $q < (2\theta - 1)/\theta$ and $q \neq 2\theta - 1$. If the mediator cannot commit to refuse mediation, any truthful equilibrium involves a probability of peace that is strictly lower than can be achieved by a mediator with commitment power.*

In a later part of the paper (Section 6) we describe and test equilibria of the direct communication game and of the human mediator game played in the lab. However, Proposition HMS, Proposition 1, and Proposition 2 are sufficient to establish the hypotheses at the heart of our experiment: the optimal mediation program is expected to yield both higher peace and higher sincerity than either of the two games.

The comparison of mediation with and without commitment is complicated by the fact that, in the absence of commitment, the revelation principle does not apply (Bester and Strausz, 2000

[12] Focusing on equilibria in undominated strategies eliminates trivial equilibria where, regardless of messages, all players' demand equals w under unmediated communication, or all players reject the mediator's recommendation under mediation. (In the lab, the frequency of subjects choosing $d = w$ in the direct communication treatment is always less than 5 percent).

and 2001).[13] Thus we consider possible lessons from the human mediator treatment more tentative. The main focus of the experiment is the relative performance of the optimal mediation program to unmediated communication.

3 Experimental parameterization and Design

Throughout the experiment we fixed $\theta = 0.7$. We studied two different parameterizations of the ex-ante frequency of H types: $q = 1/2$ and $q = 1/3$. The optimal program follows directly from Lemma 3 in HMS. The mediator's recommendations are the following:[14]

$\underline{q = 1/2}$. $r(l,l) = (0.5, 0.5)$; $r(h,l) = \{(0.7, 0.3)$ *with probability* $5/8$, $(0.5, 0.5)$ *otherwise*$\}$; $r(h,h) = \{(0.5, 0.5)$ *with probability* $1/2$, w *otherwise*$\}$. *The probability of peace is* $7/8 = 0.875$

$\underline{q = 1/3}$. $r(l,l) = (0.5, 0.5)$; $r(h,l) = \{(0.7, 0.3)$ *with probability* $3/4$, w *otherwise*$\}$; $r(h,h) = w$. *The probability of peace is* $7/9 = 0.778$.

Note that when q is low, L's temptation to lie is particularly strong because of the high probability of being matched to an L type and benefiting from the mediator's asymmetric recommendation in favor of an h message. Hence the mediator must refuse to mediate more often at lower q, with the counterintuitive conclusion that the probability of peace under optimal mediation is lower at lower q.

With $q = 1/2$, $q > 2\theta - 1$, and the optimal mediation program includes obfuscation: an H type who sent message h and receives recommendation $(0.5, 0.5)$ does not know whether the opponent sent message h or message l. Under sincerity, acceptance is preferable in the former case, and war in the latter; given the mediator's program, the H type is just indifferent and in equilibrium accepts.[15] With $q = 1/3$, on the other hand, $q < 2\theta - 1$, and in the sincere equilibrium the mediation program reveals the opponent's type: following message h, either the mediator refuses to mediate, or recommends $(0.7, 0.3)$, making clear that the opponent is L.

Experimentally, the difference makes the two parameterizations interesting. Theory tells us that it is the possibility of obfuscation that renders the mediator indispensable; but obfuscation also complicates the subjects' problem. Collecting data under both $q = 1/2$ and $q = 1/3$ allows us to study

[13] We cannot rule out that other, non-truthful equilibria, may lead to higher peace. HMS compare mediation without commitment and unmediated communication in their online appendix, and conclude that the two are equivalent if we restrict attention to standard truthful revelation mechanisms.

[14] The program depends on the pair of messages only: (h,l) is treated symmetrically to (l,h).

[15] Similarly, an L type who sent message l and receives recommendation $(0.5, 0.5)$ does not know whether the opponent sent message l or h. However, the uncertainty is less relevant for the L type, for whom accepting 0.5 is always weakly dominant.

how subjects react to the optimal mediation programs in the two cases.

We ran the experiment at Columbia's Experimental Lab for the Social Sciences (CELSS) with subjects recruited through the lab's ORSEE recruitment system (Greiner, 2015). Most subjects were undergraduate students at Columbia University and Barnard College. The experiment lasted about 90 minutes and earnings ranged from \$16 to \$37, with an average of \$28 (including a \$10 show-up fee). Experimental procedures were standard and are described in detail in the online Appendix (Section 9.5), where the instructions for one of the treatments are also reproduced.[16]

Subjects in each experimental session were exposed to a single parameterization, either $q = 1/2$ or $q = 1/3$, but to four different treatments, varying the communication and mediation protocol. Each treatment was presented as a separate part of an experimental session, consisting of multiple rounds, and instructions for each part were read just before that part began. As we describe later, with the exception of the first treatment (NC), the order of the treatments changed across sessions. To avoid decimals, the size of the resource was set to 100. We implemented the following design.

No-communication (NC)

In the no-communication treatment (NC) there was no message stage. Subjects were matched in pairs, randomly and anonymously, and independently assigned types by the computer according to q. After learning their type, each player expressed one of the feasible demands $d \in \{30, 50, 70, w\}$. If the two demands were compatible, they were satisfied; if not, the resource shrank and was shared according to the players' types. Each subject was informed of the opponent's demand and of the final outcome. Across rounds, types were reassigned and pairs rematched. We began all sessions with ten rounds of the NC treatment because their relative simplicity helped the subjects understand the game. Although those rounds were rewarded, we consider them akin to practice rounds.

Unmediated communication (UC)

The UC treatment corresponds exactly to the unmediated communication game described in the previous section. After being randomly matched in pairs and assigned a type according to q, all subjects sent their partner a message, chosen among $\{h, l, s\}$. After messages were exchanged, demands were chosen, again within the set $\{30, 50, 70, w\}$; demands were satisfied if compatible, and, if not, the resource shrank and was allocated according to players' types. As in the NC treatment, each subject was informed of the opponent's demand, and of the final outcome. In each session, we played 20 rounds of the UC treatment.

[16] The experiment was programmed in ZTree (Fischbacher, 2007).

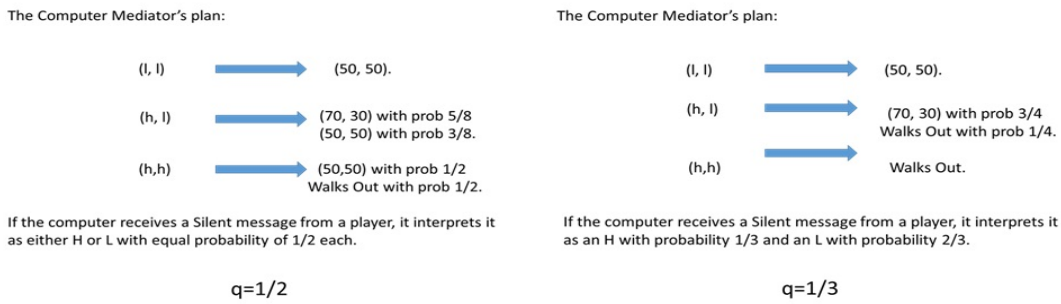

Figure 1: The Computer Mediator program.

Computer mediator (CM)

In the computer mediation treatment, we introduced the mediator, delegating the mediator's role to the computer and implementing the optimal mediation program. After having been randomly matched in pairs and assigned types, each subject sent to the computer-mediator a private message chosen among $\{h, l, s\}$. The computer then either accepted to mediate and recommended a division of the resource, or refused to mediate: $r \in \{(30, 70), (50, 50), (70, 30), w\}$. The decision was a function of the two messages, according to the optimal HMS program. The mediator's program relevant to the parameterization used in the session was projected on the lab screen during instructions and remained on the screen throughout all rounds of the treatment (Figure 1).

The instructions (and the screen) also told the subjects that the computer would interpret silence according to the prior. Unless the computer chose w, the recommendation was conveyed to each subject who then chose, separately, either to accept it or reject. If the computer issued a recommended allocation, it was implemented only if accepted by both subjects. If not, or if the computer chose w, the resource shrank and was allocated according to subjects' types. Subjects always learnt their payoff from the round. Note that if the computer mediator proposed a peaceful division, subjects could infer the opponent's message when $q = 1/3$, but not necessarily when $q = 1/2$. Each session included 20 rounds of the CM treatment.

Human mediator (HM)

In the HM treatment, in each round subjects were randomly matched in groups of three; two players and one mediator. The round proceeded following the mediation game rules, but without constraining

the mediator to any specific program and without communicating any such program to the players. After privately learning their type, players 1 and 2 each sent a confidential message $m \in \{h, l, s\}$ to the mediator. The mediator knew q, but had no additional information. Upon receiving the messages, the mediator issued a recommendation $r \in \{(30, 70), (50, 50), (70, 30), w\}$. Unless $r = w$, each of the two players, independently, could either accept the recommendation or reject it. If both accepted, the recommendation was implemented; if not or if $r = w$, conflict followed, the resource shrank to 70 and was allocated according to the players' types. All subjects learnt the outcome of the game–whether the recommendation was made and accepted, and in all cases their payoffs; but did not learn the opponent's message and, unless there was conflict, the opponent's type.

Note that the mediator lacks commitment power. We are interested in exploring the impact of the lack of commitment, but worried that with the mediator having no incentive ever to refuse mediation, subjects would converge to a trivial equilibrium with all messages pooled at h.[17] Thus we rewarded the mediator according to the following schedule: the mediator earned 60 if a recommendation was made and accepted, 20 if the recommendation was made but was rejected, and 40 if the mediator refused to mediate. The numerical values for the mediator's payoffs were kept constant across the two parameterizations. The mediator's payoff schedule introduced the incentive to refuse to mediate when mediation was likely to fail; if rejection is most probable when the two players have both reported message h, the indirect effect is to discipline the L type by checking its temptation to send message h. Proposition 2 continues to apply to our HM game.

As under all previous treatments, players always learnt the opponent's type in case of conflict, but not if a peaceful division was achieved. In addition, since the mediator's program was not announced to the players and most probably was not consistent across mediators, they could not deduce the opponent's message from the mediator's recommendation with any confidence.

At each round, the three subjects were matched randomly, but under the constraint that all subjects played the role of mediator for an equal number of rounds. In each session, subjects played 30 rounds of the HM treatment, 10 rounds as mediator and 20 as players.

Because our main focus is on the comparison between the UC and the CM treatments, we varied the order of treatments so as to treat UC and CM symmetrically. We ran 12 experimental sessions, each with 12 subjects, with the following experimental design:

[17] It is not difficult to see that such an equilibrium exists (the mediator ignores the messages and always recommends $(30, 70)$; all types accept 70, L accepts 30, and the probability of peace is $(1 - q)$).

parameterization and order of treatments

Sessions	q	Order
s1-s3	1/2	1: NC, UC, HM, CM
s4-s6	1/3	1: NC, UC, HM, CM
s7-s9	1/2	2: NC, CM, HM, UC
s10-s12	1/3	2: NC, CM, HM, UC

Number of subjects, groups and rounds per session

Order	# Subjects	# Groups per Treatment	# Rounds	Groups × Rounds
1: NC, UC, HM, CM	12×3	6,6,4,6	10,20,30,20	60,120,120,120
2: NC, CM, HM, UC	12×3	6,6,4,6	10,20,30,20	60,120,120,120

Because we always ordered NC first and, as mentioned, treated it as a practice treatment, we do not compare its results to the other treatments and do not discuss it in the text. For completeness, we describe the NC data as well as the equilibria of the NC game in the online Appendix (Section 9.4.4).

4 Experimental Results

When comparing unmediated communication and mediation, the theory makes two broad qualitative predictions. The optimal mediation program can lead to: (1) more sincerity, and to (2) more frequent agreement. The CM treatment implements the program that can support the best equilibrium, but whether the results are observed in the lab depends on the predictive power of such an equilibrium. Whether the predictions extend to the human mediator treatment, with inexperienced mediators and inexperienced players, is an interesting but still more speculative question.

We begin by reporting our results on sincerity.

4.1 Sincerity

The two panels of Figure 2 report the frequencies of different messages in the two parameterizations, $q = 1/2$ and $q = 1/3$, for the three treatments, UC, HM and CM. In each panel, the H type's messages are reported on the left, and the L type's messages on the right. The data are aggregated

over all sessions and both orders of treatments. Confidence intervals are calculated from standard errors clustered at the individual level.

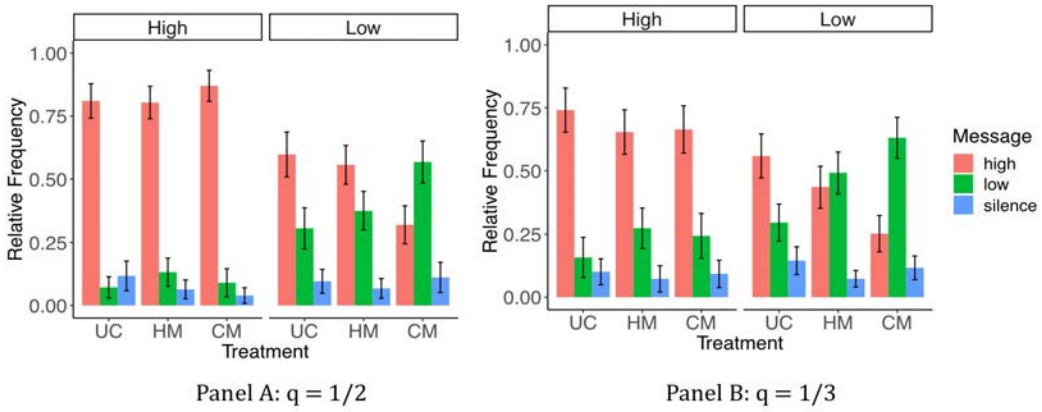

Figure 2: Messages by type and treatment.

The figure makes clear a number of regularities. First, although we never see full sincerity, H types send message h with high frequency in all treatments and for both parameterizations. In all treatments more than 80 percent of all H types send message h when $q = 1/2$; more than 65 percent do so when $q = 1/3$. There is no detectable treatment effect. Note that H types' truthfulness should not be taken for granted: even under CM, the incentive compatibility constraints of the H types are binding when taking into account the possibility of double deviation (an untruthful message followed by rejection of the mediator's recommendation). Second, there is less sincerity but a clear treatment effect for L types: the frequency of l messages from L subjects goes from 31 percent in UC to 57 percent in CM if $q = 1/2$, and from 30 to 64 percent if $q = 1/3$. The difference is significant both quantitatively and statistically. The HM treatment too sees higher frequency of sincere l messages, relative to UC: 38 percent with $q = 1/2$ and 49 with $q = 1/3$. Third, the option of sending a silent message is used relatively little: it is always less than 15 percent of messages sent by either type.[18]

As shown in the first column of Table 1, a linear probability model confirms what the figures show.[19] L types are less sincere than H types, and for L types treatment effects are present and

[18] Using data from Order 1 and Order 2, our design allows us to compare the UC and CM treatment between subjects, when the two treatments are run in rounds 11-30 in a session, and thus on subjects with identical experience. Figure 14 in the online Appendix (Section 9.4.1) replicates Figure 2 and shows the same regularities.

[19] Results are unchanged under a probit model. We report all regression results in the paper as estimated from a

	Dependent variable:	
	Sincerity	Silence
	(1)	(2)
HM Treatment	-0.039	-0.042^{**}
	(0.028)	(0.020)
CM Treatment	0.005	-0.049^{**}
	(0.034)	(0.021)
Order 2	0.009	-0.069^{***}
	(0.040)	(0.025)
$q = 1/2$	0.142^{***}	-0.015
	(0.042)	(0.026)
L-type	-0.354^{***}	-0.029
	(0.082)	(0.039)
Round	0.002^{***}	-0.001^{***}
	(0.001)	(0.0004)
HM treatment $\times$ L-type	0.181^{***}	-0.012
	(0.042)	(0.021)
CM treatment $\times$ L-type	0.298^{***}	0.038
	(0.055)	(0.028)
Order 2 $\times$ L-type	-0.025	0.018
	(0.062)	(0.021)
$q = 1/2 \times L$-type	-0.198^{***}	-0.005
	(0.063)	(0.022)
Round $\times$ L-type	-0.0002	0.001
	(0.001)	(0.001)
Constant	0.611^{***}	0.212^{***}
	(0.057)	(0.041)
Observations	8,640	8,640
R^2	0.158	0.024
Adjusted R^2	0.157	0.022
Residual Std. Error (df = 8628)	0.453	0.288
Note:	*p<0.1; **p<0.05; ***p<0.01	

The excluded category in the regression is H in treatment UC with $q = 1/3$ under Order 1. Standard errors are clustered at the individual level.

Table 1: Sincerity and Silence.

significant: sincerity is lowest under UC, intermediate under HM and highest under CM. In addition, H types, but not L types, are more sincere when $q = 1/2$, and both types learn to become more sincere with experience in the session, but the effect is very small.

The data show the extent to which subjects assigned different types report their type sincerely. We should keep in mind, however, that sincerity does not map directly into the information conveyed. How much a message moves the posterior probability of a given type relative to the prior depends on the use of the message by both types.[20] In the online Appendix (Section 9.4.3), we report Kullback Leibler (KL) measures applied to our data. The measures are instructive, but it remains true that the treatment conveying most information is CM, for both messages and in both parameterizations.

Finally, our experimental design allows subjects to send a silent message. As Figure 2 shows, silent messages are used sparingly–the highest frequency is 14 percent by L types in UC, with $q = 1/3$. Estimation of a linear probability model shows that silence is used more frequently in UC, when communicating directly, and its frequency declines with experience (Table 1, column 2). The reduced use of silence under Order 2 is probably due to the early experience with the mediation treatments (recall that UC is the last treatment under Order 2).

4.2 Peace

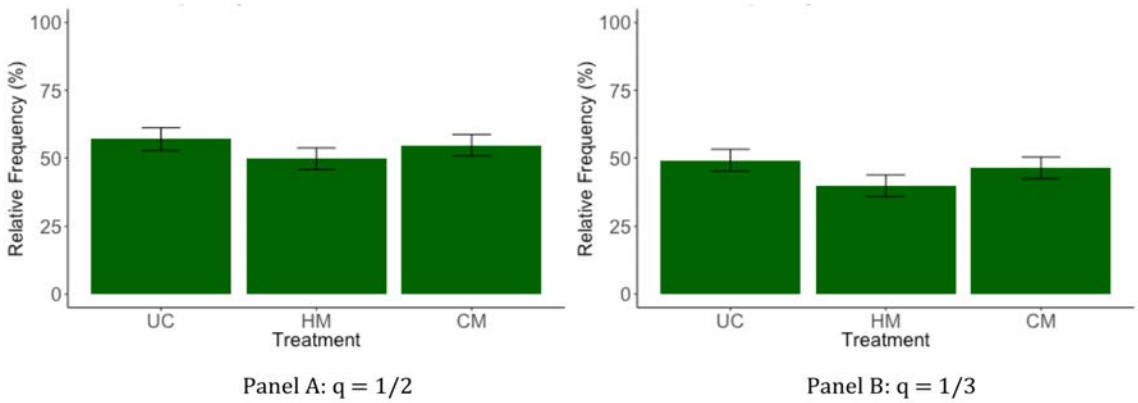

Figure 3: Frequency of peace.

linear probability model; in all cases we have verified that qualitative results are unchanged under probit.

[20] To take an extreme example, suppose H types are fully sincere and always send message h; yet, if L types always lie, no information is conveyed by an h message because *all* types always send message h.

	Dependent variable:	
	Peace	
	(1)	(2)
HM Treatment	-0.082^{***}	-0.085^{***}
	(0.020)	(0.018)
CM Treatment	-0.025	-0.019
	(0.020)	(0.018)
Order 2	0.005	0.012
	(0.018)	(0.016)
$q = 1/2$	0.087^{***}	0.186^{***}
	(0.018)	(0.016)
H-L pair		0.293^{***}
		(0.018)
L-L pair		0.606^{***}
		(0.018)
Round	0.001^{***}	0.001^{***}
	(0.0004)	(0.0003)
Constant	0.435^{***}	0.042
	(0.026)	(0.026)
Observations	4,320	4,320
R^2	0.015	0.191
Adjusted R^2	0.013	0.190
Residual Std. Error	0.497 (df = 4314)	0.450 (df = 4312)
Note:		*p<0.1; **p<0.05; ***p<0.01

The default treatment is UC, Order 1, $q = 1/3$, and when looking at different pair types, the default pair is H-H. Standard errors are clustered by pairs of subjects.

Table 2: Peace.

Figure 3 reports the frequency of peace across treatments, for the two parameterizations, as well as 95 percent confidence intervals (with standard errors clustered at the pair of subjects level). Whether $q = 1/2$ or $q = 1/3$, there is no significant difference across treatments. In both cases, treatment HM results in least frequent peace and UC in most frequent peace, but the effects are small.[21] In both cases, the highest theoretical frequency under CM (87 percent with $q = 1/2$, and 78 percent with $q = 1/3$) is not within the confidence interval. On the other hand, the difference between the two parameterizations is in the direction the theory predicts, with higher peace in all treatments under $q = 1/2$, a finding we study in more detail and confirm in the online Appendix (Section 9.4.2).

The estimation of a simple linear model of the frequency of peace, isolating treatment, order and parameter effects, qualifies the results slightly but does not change the main message. We report the results in Table 2 below, where we also add the round number, to control for learning, and the pair types. As expected, peace is highest between $L-L$ pairs, and lowest between $H-H$ pairs, it is higher under $q = 1/2$, and increases significantly but very little over time. Across treatments, UC and CM are comparable, while peace is significantly lower under HM.[22]

5 Optimal Mediation Induces More Sincerity but not More Peace. Why?

Theory gives us precise hypotheses for two of the treatments, UC and CM. What can we learn from comparing the experimental results to the theoretical predictions in these two cases? The lesson from the data is unambiguous: because of the messages by L types, sincerity is higher, and the messages more informative, under CM. But peace is not. We can represent both observations in a single graph. Call τ_T the frequency with which a type T player sends a truthful message, and σ_T the frequency with which the player is silent. Recall that in CM the computer interprets silent messages according to the prior. Thus we define $\widehat{\tau}_L = \tau_L + (1-q)\sigma_L$ as the frequency of all messages sent by L subjects that are read as l by the computer, and $\widehat{\tau}_H = \tau_H + q\sigma_H$ is the frequency of all messages sent by H subjects that are read as h by the computer.

[21] Ordering the numbers as $\{UC, HM, CM\}$, the frequencies of peace in the data are: $\{0.57, 0.50, 0.55\}$ if $q = 1/2$, and $\{0.49, 0.40, 0.46\}$ if $q = 1/3$.

[22] In the online Appendix (Section 9.4.1), Figure 15 reports the frequency of peace in CM and UC between subjects, when both treatments are played second in the session (the conclusion is identical). We also report in the online Appendix (Section 9.4.2) the regression results with the full set of interaction terms. We find that the low performance of the HM treatment is driven by the unusually low frequency of agreement when one or, especially, when both players are of type L.

We construct descriptive figures of the data by imputing the same interpretation of silence in UC, i.e. by supposing that subjects too interpret silence according to the prior.[23]

For each experimental session, Figure 4 reports $\widehat{\tau}_L$ on the horizontal axis, $\widehat{\tau}_H$ on the depth axis, and the frequency of peace on the vertical axis. Panel A refers to $q = 1/2$, panel B to $q = 1/3$. Each sphere corresponds to a session; yellow spheres report results for UC treatments, and red spheres for CM treatments. The two green cubes correspond to the theoretical equilibria with highest peace in the two treatments[24] (the green cube centered among the yellow spheres refers to UC; the green cube corresponding to $\widehat{\tau}_L = 1$ and $\widehat{\tau}_H = 1$ represents the HMS equilibrium in CM). As shown earlier, the two treatments on average yield similar values for $\widehat{\tau}_H$. Here, yellow and red spheres align similarly along the depth axis (not easily readable in the figure), but are clearly differentiated along the horizontal axis, and the orientation of the figures highlights the two clusters, almost fully distinct, with lower $\widehat{\tau}_L$ values for UC, and higher $\widehat{\tau}_L$ values for CM. However, the spheres are not organized by color on the vertical axis–the frequency of peace. There is no systematic variation between the two treatments.

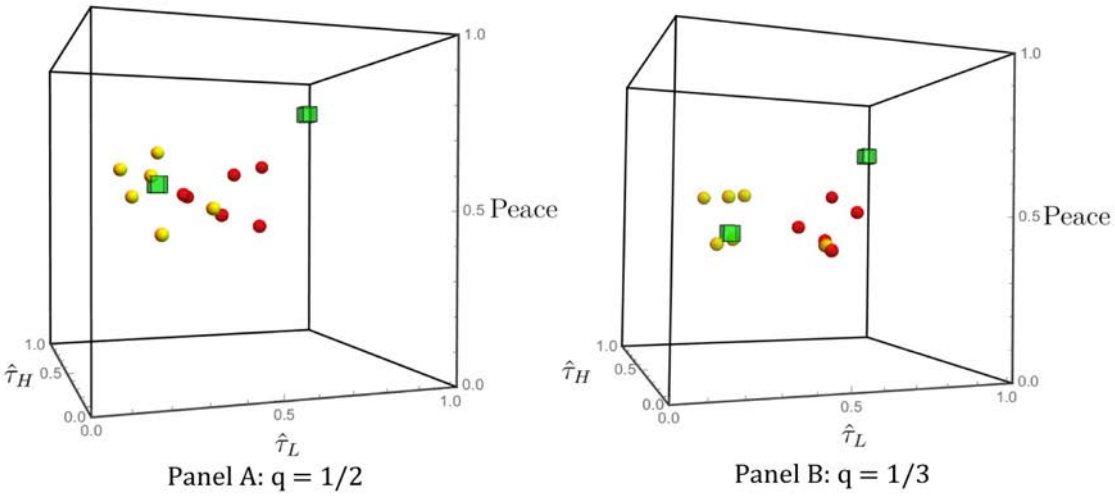

Figure 4: Sincerity and peace in UC and CM sessions.

Why wasn't the promise of optimal mediation realized in the data? Figure 5 gives some indications

[23] Recall that the frequency of silence is low and comparable both across types and across the two treatments, in both parameterizations (Figure 2).

[24] For UC, the green cube corresponds to the equilibrium with highest peace among the equilibria characterized in Section 6.2.

of where the problems lie. The figure plots, for each parameterization, the causes of war under CM in the data. The orange columns correspond to the computer's refusals to mediate, either in the data (lighter orange), or if all subjects had been sincere (darker orange); green columns indicate rejections of the computer's offer by H types, and blue columns by L types, organized according to the offer.[25] In the optimal mediation equilibrium, all messages are sincere, all offers are accepted, and war only follows from the mediator's refusal to mediate. In the data, not all messages are sincere and not all recommendations are accepted, and the figure reflects both types of deviations.

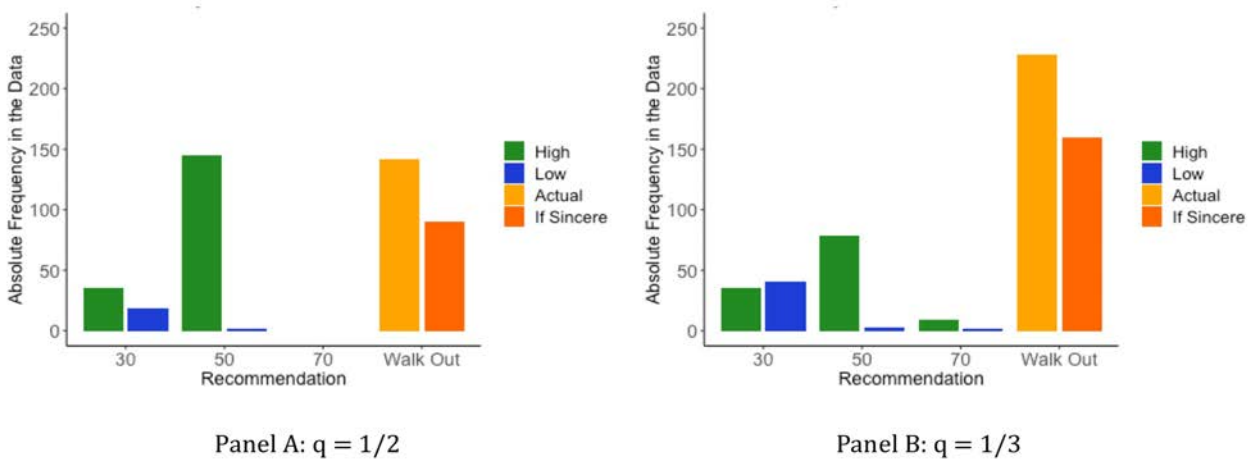

Figure 5: Causes of war.

With both parameterizations, dominated actions (L rejecting 50, either type rejecting 70) are rare. When $q = 1/2$, excess war has two main causes. The first is the lack of full sincerity of L types, reflected in the higher frequency of refusals to mediate. The second, more striking, is the high number of rejections of offers of 50 by H types: sincere H types rejected more than one third of all 50 offers they received. The subtlety of the obfuscation appears not to work in the lab.

When $q = 1/3$ as well, the two dominant causes of war are L's lack of full sincerity, reflected in the high frequency of refusals to mediate, and H's refusals of offers of 50. But with $q = 1/3$, H types are not recommended 50 if they are sincere. The rejections we see in the data arise from H types' frequency of lies (see Figure 2).

[25] The figures report individual rejections of offers. Because a single rejection is sufficient to trigger war, there can be some double counting: two individual rejections can amount to a single offer being turned down.

The lack of full sincerity in the lab is hardly surprising; what is surprising is the low success of mediation in achieving peace. One possible explanation is that the CM treatment has multiple equilibria. The HMS equilibrium is the equilibrium with highest peace, but, given the mediation program, how sensitive are the other equilibria to less than full truthfulness?

5.1 Multiple equilibria under computer mediation

Keeping fixed the mediator's program, we study the equilibria of the CM treatment in undominated strategies.[26] We concentrate on equilibria where, regardless of message: (i) all players accept 70; (ii) L players always accept 50; (iii) H players always reject 30. Denoting by Tm a player of type T who sent message m, what remains to be determined are the acceptance strategies of Hh and Hl players offered 50, and of Ll players offered 30, as well as the first stage message strategies for both types. We simplify notation by denoting by α_m the probability of an Hm type accepting 50, and by β the probability of an Ll type accepting 30 (the offer of 30 can only follow an l message). As before, we denote by $\widehat{\tau}_T$ the probability that type $T \in \{H, L\}$ is truthful (corrected for silence).

We report the full set of equilibria in the Appendix (Section 8.2); here we concentrate on equilibria that do not contradict grossly the experimental data. In particular, in the data, having sent message l, L types accept 30 more than 89 percent of the times if $q = 1/2$, and 80 percent of the times if $q = 1/3$. In line with this observation, we focus here on equilibria with $\beta = 1$. By selecting such equilibria, we also rule out equilibria where $\widehat{\tau}_H < \widehat{\tau}_L$, in clear contradiction to our data. The equilibria are reported in Table 3 and represented graphically in Figure 6.[27]

$q = 1/2$	$q = 1/3$
$\alpha_h = 1, \beta = 1, \widehat{\tau}_L = 1, \widehat{\tau}_H = 1$	$\beta = 1, \widehat{\tau}_L = 1, \widehat{\tau}_H = 1$
$\alpha_h = 0, \beta = 1, \widehat{\tau}_L = 1, \widehat{\tau}_H = 1$	$\alpha_l = 0, \beta = 1, \widehat{\tau}_L \in (0,1), \widehat{\tau}_H = 1/3 + (2/3)\widehat{\tau}_L$
$\alpha_l = 0, \alpha_h = 0, \widehat{\tau}_L = 0, \widehat{\tau}_H \in [1/6, 4/15]$	
$\alpha_l = 0, \alpha_h = 0, \beta = 1, \widehat{\tau}_L \in (0,1), \widehat{\tau}_H = 4/15 + (6/15)\widehat{\tau}_L$	
$\alpha_l = 0, \alpha_h = 0, \beta = 1, \widehat{\tau}_L = 1, \widehat{\tau}_H \in [2/3, 1)$	

[26] Because our focus is understanding the experimental results, we characterize the equilibria for the specific parameter values used in the experiment. The analysis generalizes to arbitrary θ and q, keeping in mind that $q = 1/2$ corresponds to $q > (2\theta - 1)$ and $q = 1/3$ to $q < (2\theta - 1)$, the mediation program corresponds to Lemma 3 in HMS, and we maintain the assumptions $q < (2\theta - 1)/\theta$ and $\theta/2 > (1 - \theta)$.

[27] With $q = 1/2$, the equilibria with $\widehat{\tau}_L = 0$ are supported by the off-equilibrium belief $\beta = 1$.

Table 3. Equilibria under CM

For both parameterizations, the first equilibrium in Table 3 is the HMS equilibrium, identified by the green cubes in Figure 6; the other equilibria correspond to the red lines. For both $q = 1/2$ and $q = 1/3$, there are equilibria supporting a large range of peace probabilities, between 0 and the highest frequency, corresponding to the HMS equilibrium (albeit with a discontinuous jump for $q = 1/2$). Similarly, for both values of q, there are equilibria spanning any frequency of truthfulness, from 0 to 1, for L's and almost as large a range for H's. This then is our first observation: keeping fixed the mediator's program, equilibrium behavior under CM is compatible with a large range of messages and outcomes.

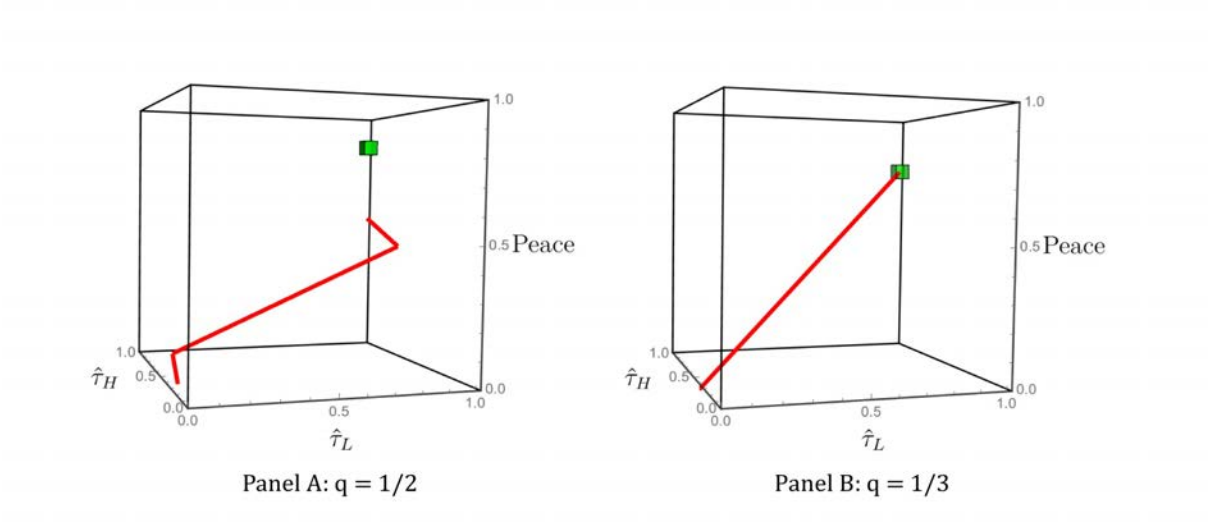

Figure 6: Equilibria under CM.

Beyond depicting such wide variation, the most striking feature of the figure is the discontinuity in the locus of equilibria under $q = 1/2$. The mediation program uses obfuscation, and the HMS equilibrium entails $\alpha_h = 1$: sincere H types accept the $(50, 50)$ offer with probability 1. The figure shows that the equilibrium is fragile: if there is *any* deviation in the messages from full sincerity by either type, including any use of the silent message, the peace probability falls discontinuously.

The discontinuity does not depend on the specific parameters used in the experiment but applies over the whole parameter region for which obfuscation is part of the optimal mediation program. And because it is obfuscation that makes the HMS equilibrium superior to any equilibrium of the direct

communication game, the observation is of interest beyond the specific experiment. We phrase it in the following proposition for generic parameter values in the appropriate range. In line with the HMS model, Proposition 3 ignores the option of silent messages.[28] (We include the possibility of silence in the Appendix, where the result in the proposition is part of the equilibria characterization, specialized to the experimental design and parameters).

The relevant restrictions are $(2\theta - 1) < q < (2\theta - 1)/\theta$, the range of parameter values for which obfuscation is optimal. Following Lemma 3 in HMS, the optimal mediation program is then the following: $r(l,l) = (1/2, 1/2)$; $r(h,l) = \{(1/2, 1/2)$ with probability q_M and $(\theta, 1-\theta)$ otherwise$\}$; $r(h,h) = \{(1/2, 1/2)$ with probability q_H and w otherwise$\}$ where:

$$q_M = \left(\frac{1-\theta}{2\theta - 1}\right)\left(\frac{1+q-2\theta}{\theta - q}\right) \tag{1}$$
$$q_H = \left(\frac{1-q}{q}\right)\left(\frac{1+q-2\theta}{\theta - q}\right).$$

As above, we use α_h (α_l) to denote the probability that type Hh (Hl) accepts 1/2. The following proposition holds:

Proposition 3. *Suppose* $(2\theta - 1) < q < (2\theta - 1)/\theta$. *Then*: *(i)* $\alpha_h = 1 \Longrightarrow \{\tau_H = 1, \tau_L = 1\}$. *(ii) If* $\tau_H < 1$ or $\tau_L < 1$, *then* $\alpha_h = 0$. *(iii)* $\{\tau_H = 1, \tau_L = 1\} \nRightarrow \alpha_h = 1$.

Proof. Call $\Delta_{Hh}(1/2)$ the expected differential gain from accepting rather than rejecting 1/2 for player i, an H player who sent message h. Player i's opponent is indexed by j, and we indicate by $\Pr(T_j)$ the probability that j is a type T and by $\Pr(Tm_j)$ the probability that j is a type T who sent message m. Since all L types always accept 1/2, it is not difficult to see that:

$$\Delta_{Hh}(1/2) = (1/2 - \theta/2)[\Pr(Hh_j \mid (1/2,1/2), h_i)\alpha_h + \Pr(Hl_j \mid (1/2,1/2), h_i)\alpha_l] +$$
$$(1/2 - \theta)\Pr(L_j \mid (1/2,1/2), h_i)$$

[28] HMS allow for some probability $p < 1/2$ that L prevails in case of conflict. Since we have set $p = 0$ throughout, we do not reintroduce it here, but we have verified that the discontinuity is robust to the generalization.

Solving the probabilities:

$$\Pr(Hh_j \mid (1/2,1/2), h_i) = \frac{\Pr(Hh_j \text{ and } (1/2,1/2)|h_i)}{\Pr((1/2,1/2)|h_i)}$$
$$= \frac{q_H \tau_H q}{q_H[\tau_H q + (1-\tau_L)(1-q)] + q_M[(1-\tau_H)q + \tau_L(1-q)]}$$

$$\Pr(Hl_j \mid (1/2,1/2), h_i) = \frac{q_M(1-\tau_H)q}{q_H[\tau_H q + (1-\tau_L)(1-q)] + q_M[(1-\tau_H)q + \tau_L(1-q)]}$$

and:

$$\Pr(L_j|(1/2,1/2), h_i) = \frac{q_M\tau_L(1-q) + q_H(1-\tau_L)(1-q)}{q_H[\tau_H q + (1-\tau_L)(1-q)] + q_M[(1-\tau_H)q + \tau_L(1-q)]}.$$

Substituting (1) and simplifying, $\alpha_h > 0$ requires $\Delta_{Hh}(1/2) \geq 0$ or:

$$(1-q)\tau_H\alpha_h + \frac{(1-\theta)q}{(2\theta-1)}(1-\tau_H)\alpha_l \geq (1-q)\tau_L + \frac{(2\theta-1)(1-q)^2}{(1-\theta)q}(1-\tau_L). \tag{2}$$

The left-hand side of (2) is always weakly increasing in α_h and α_l, and maximal at $\alpha_h = \alpha_l = 1$ and $\tau_H = 1$, while the right-hand side is minimal at $\tau_L = 1$. Hence the condition is mostly likely to be satisfied at these values, at which it simplifies to $(1-q) = (1-q)$, holding with equality. Thus if $\alpha_h > 0$, then $\alpha_h = 1$, $\tau_H = 1$, $\tau_L = 1$. If either $\tau_H < 1$ or $\tau_L < 1$, then $\alpha_h = 0$. In addition, even at $\tau_H = 1$, $\tau_L = 1$, a second equilibrium exists with $\alpha_h = 0$: full sincerity is necessary but not sufficient for $\alpha_h = 1$. □

Keeping the mediation program constant, any expected deviation from full sincerity by others induces the Hh type to *always* reject 1/2–$\alpha_h = 0$. In fact, even with full sincerity a second equilibrium exists where $\alpha_h = 0$. The intuition is straightforward: when offered 1/2, H's best option is to accept if the opponent is H and reject if the opponent is L, conditional on the opponent accepting. If other H's are expected to reject, always rejecting is a best response, even if all are sincere. And even if other H's are expected to accept, rejecting is a best response if the posterior probability that the opponent is L, conditional on the mediator's recommendation, is high enough–and simple calculations show this must indeed be the case for any deviation from full truthfulness by either type.

Proposition 3 is very relevant for a lab experiment and possibly for actual applications of mediation plans: expecting some lying with positive probability seems inevitable. The proposition tells us that, for the relevant range of parameter values, no peace probability in the neighborhood of the HMS equilibrium should then be expected. Under the optimal mediation program, the ex-post participation

constraint of an H type offered 1/2 is binding; any deviation leads to the violation of the constraint.[29] Yet, what is interesting is that the discrete jump in the expected frequency of peace is limited to the mediation program that exploits obfuscation. As shown in Figure 6, panel B, in the absence of obfuscation under $q = 1/3$ there is no discontinuity in the locus of equilibria around the full sincerity point: a small probability of untruthful messages leads to a lower probability of peace, but the equilibrium analysis shows that compliance of sincere types with the mediator's recommendations is not affected. This is true if $q = 1/3$, or more broadly $q < (2\theta - 1)$, and the optimal mediation program does not include obfuscation. It is also true if $q > (2\theta - 1)$ and the mediation program is optimized under the constrain of no obfuscation. The reason is that in the absence of obfuscation the ex post participation constraints for a sincere H type offered 1/2 and a sincere L type offered $(1 - \theta)$ are slack, and remain slack in the presence of lies; the ex post participation constraints for a sincere H type offered θ is binding under full sincerity and remains binding along the equilibrium locus in the presence of lies, but acceptance is weakly dominant.[30]

5.2 Sincerity and peace: data v/s equilibrium predictions

Do the multiple equilibria explain the relatively poor performance of mediation in the CM treatment? Figure 7 below superimposes the data, aggregated by session, to the equilibria in Figure 6. As before, the horizontal axes measures $\widehat{\tau}_L$ and $\widehat{\tau}_H$, (sincerity, corrected for silence), and the vertical axis the frequency of peace (or the ex ante probability of peace in the theoretical model). The equilibria correspond to the red lines, with the equilibrium with highest expected peace–the HMS equilibrium– denoted by the green cube. The data are represented by red spheres.

The 3D figures allow us to represent directly the three variables at the heart of the mechanism,

[29] Interestingly, the obfuscation equilibrium with $q = 1/2$ is trembling-hand perfect. Perfection arises because of the latitude left to the specification of the error processes. As long as the probability of trembles in L's acceptance strategies is high enough, relative to the probability of the other trembles, the equilibrium is perfect. (See the discussion in the Appendix (Section 8.3)).

[30] Without obfuscation, subjects learn their opponent's message from the mediator's recommendations. Hence for example $\Delta_{Hh}(1/2)$ becomes:

$$\Delta_{Hh}(1/2) = (1/2 - \theta/2)\frac{\tau_H q}{\tau_H q + (1-\tau_L)(1-q)}\alpha_h + (1/2 - \theta)\frac{(1-\tau_L)(1-q)}{\tau_H q + (1-\tau_L)(1-q)}$$

Thus, $\Delta_{Hh}(1/2) \geq 0$ if:

$$(1-\theta)\tau_H q \alpha_h \geq (2\theta - 1)(1-\tau_L)(1-q)$$

Whether $q > (2\theta - 1)$ or $q < (2\theta - 1)$, there always are $\underline{\tau_L} < 1$ and $\underline{\tau_H} < 1$ such that accepting 1/2 is superior to rejecting it if $\alpha_h = 1$, $\tau_L \in (\underline{\tau_L}, 1)$, and $\tau_H \in (\underline{\tau_H}, 1)$. With $q = 1/2$ and $\theta = 0.7$, the optimal mediation program in the absence of obfuscation corresponds to: $r(l,l) = (1/2, 1/2)$; $r(h,l) = (0.7, 0.3)$; $r(h,h) = (1/2, 1/2)$ with probability 1/5, and w otherwise. The expected frequency of peace is 0.8 (v/s 0.875 with obfuscation).

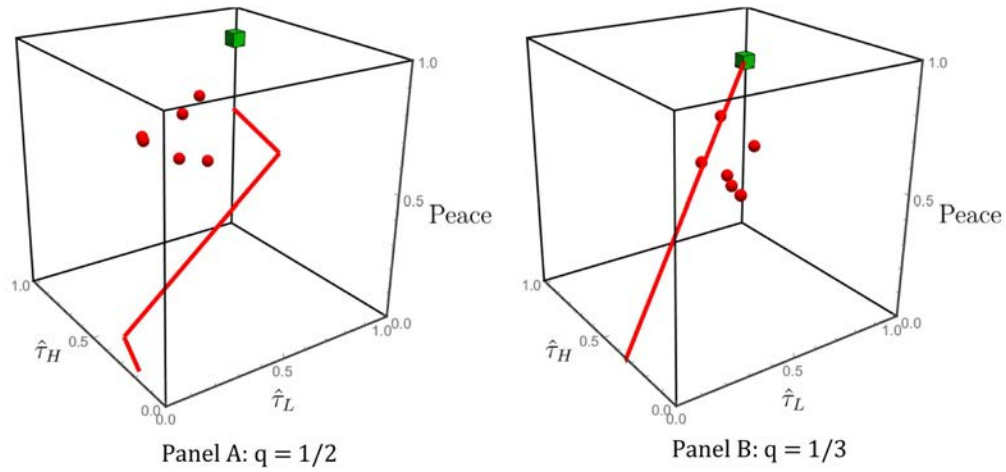

Figure 7: CM: Data and equilibria.

sincerity by either type and peace. As we already know, in both parameterizations and all sessions, all three variables fall short of the HMS equilibrium. Figure 7 shows that, relative to the other equilibria, the deviations for the two parameterizations go in opposite directions: with $q = 1/2$, holding $\widehat{\tau}_L$ fixed at the experimental values, the data have more frequent peace and more sincere H types than the theory predicts; with $q = 1/3$, two sessions sit almost exactly on the equilibrium line; in the remaining four, holding $\widehat{\tau}_L$ fixed at the experimental values, the data have less peace and less sincerity from H types than the corresponding theoretical equilibria.

With $q = 1/3$, untruthful messages by H's are followed by recommendations of either $(50, 50)$ or $(30, 70)$, which are then typically rejected. Had those messages been sincere, some would have been followed by recommendations of $(70, 30)$, which could have resulted in peace.[31] But note that peace would have occurred only if the opponent sent message l[32] and accepted 30, that is, if the opponent was a sincere L. And in this case the payoff to the H type would be 70, whether from peace or from war. In other words, with $q = 1/3$, H's payoff from sincerity and compliance is identical to the payoff from message l (or s), and then the rejection of any recommendation of either 50 or 30. The loss of efficiency comes at no cost to the H player. In such a situation, it is not implausible that other

[31] Both types Hl and Hs reject 50 more than 80 percent of the times, and reject 30 100 percent of the times. Types Ll and Ls accept 30 more than 80 percent and 60 percent of the times, respectively.

[32] Recall that the computer mediator always walks out after messages (h, h).

considerations may play a role. For example, the desire to maintain control over triggering conflict, as opposed to having it imposed by the computer mediator, could explain the relatively high frequency of H's lies.

With $q = 1/2$, peace is instead higher than equilibrium predicts. In all equilibria with less than perfect truthfulness H types never accept 50. In the data, aggregating over all sessions, the frequency of acceptances is just below 60 percent, with high dispersion across subjects. Figure 8 shows the disparity in individual strategies. The vertical axis is the frequency of acceptance of 50 by an H type; the x and y-axes are $\widehat{\tau}_L$ and $\widehat{\tau}_H$, and the red line at the bottom is the equilibrium line.[33] There are 72 subjects in all, and each is represented in the figure by the subject's frequency of truthfulness (as an H and as an L), and acceptance of 50 when H; the volume of each sphere is proportional to the number of individual subjects the sphere represents. The green sphere corresponds to the 5 subjects who always played (1,1,1)–the HMS equilibrium strategies. There are 4 subjects at (1,1,0) (also an equilibrium point); the overwhelming majority of the rest are individual subjects. The figure shows clearly that only a few subjects reject the offer consistently.

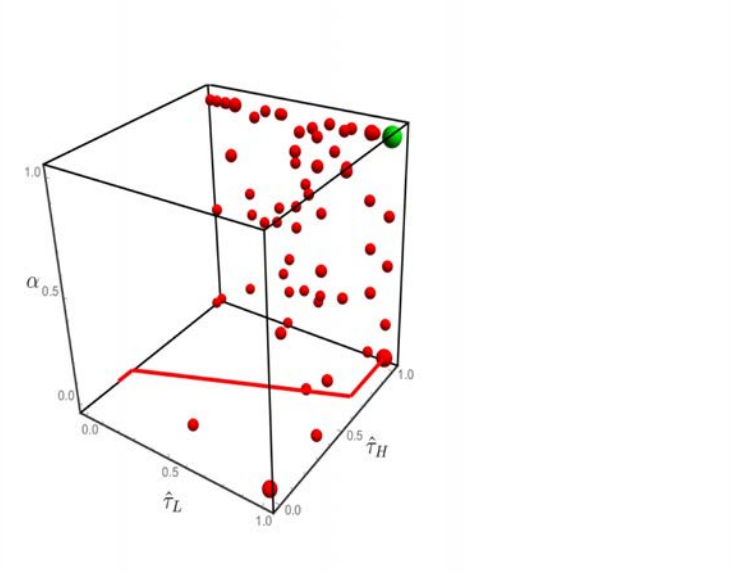

Figure 8: Individual subjects; $q = 1/2$.

Why are H types accepting 50, against the theory's predictions? Two explanations seem plausible. First, subjects could be risk averse. The optimal mediation program would differ under risk aversion, but we can still ask how risk averse subjects would respond to the program implemented by the

[33]The figure does not distinguish on the basis of the subject's message, but close to 90 percent of all messages were h.

computer mediator. Rejecting the mediator's recommendation increases uncertainty, and indeed risk aversion can induce a sincere H type to accept 50. Proposition 3 does not hold under risk aversion.[34]

We did not elicit measures of risk aversion, but we can deduce them from subjects' behavior in other treatments of the experiment–a methodology with the advantage of not disturbing the experiment or creating experimenter demand effects. Recall that we started each session with 10 rounds of direct demands without communication (NC). In the NC treatment, demanding 30 is dominated for an H (and we observe it only once, out of 356 demands) but is not dominated and is the minimum risk strategy for an L. Under $q = 1/2$, L types demand 30 30 percent of the times; for reference, the unique equilibrium of the NC game under risk neutrality has L types demanding 30 with more than 70 percent probability (see the online Appendix (Section 9.4.4)). Across subjects, the correlation between the frequency of accepting 50 when Hh in CM and demanding 30 when L in NC is negative: $\widehat{\rho} = -0.39$ (with 95 percent $CI = [-0.58, -0.17]$). In the HM treatment, recall that refusing to mediate guarantees the mediator a riskless payoff of 40–we expect risk averse mediators to walk out with high frequency. Under $q = 1/2$, following messages (h, h), human mediators walk out 36 percent of the times; as we describe in Section 6.1, the HM game has many equilibria, but in the equilibrium that best explains the data, the corresponding frequency under risk neutrality is in fact higher (0.535). Across subjects, we find no correlation between the frequency of accepting 50 when Hh in CM and walking out after (h, h) in HM: $\widehat{\rho} = -0.13$ (with $CI = [-0.36, 0.11]$). All together, these numbers do not make a case for risk aversion.

A second possible explanation for the behavior we observe is that the actions chosen by the subjects come at little cost. With the theory predicting that an H will reject any offer of 50 with probability 1, any noise will result in more acceptances and more peace. If the cost is small, some noise in behavior is to be expected. Given the behavior of others, how far are subjects from best responding? We address this question in the next section.

[34] Under the CM program,

$$\Delta_{Hh}(50) = [u(50) - u(35)][4\tau_H\alpha_h + 3(1-\tau_H)\alpha_l] + [u(50) - u(70)](4 - \tau_L)$$

It is not difficult to verify that the constraint now has slack at full sincerity and $\Delta_{Hh}(50) \geq 0$ is possible under some lying. Note however that the truthful equilibrium where H types always reject 50 ($\alpha_h = 0, \alpha_l = 0, \tau_H = 1, \tau_L = 1$) continues to exist.

5.3 Neighborhood of best responses

Dominated actions are rare in the data. If we ignore them, each player of given type faces two decisions: the message, $\widehat{\tau}_H$ if H and $\widehat{\tau}_L$ if L, and the acceptance of 50 if H, α, and of 30 if L, β.[35] For each session, we calculated the average strategies played by others in a session. We then calculated the expected payoff of an H type as a function of $\widehat{\tau}_H$ and α, and correspondingly of an L type as a function of $\widehat{\tau}_L$ and β. Given type, the differences across subjects and sessions were minor, and our findings can be summarized in the figures below, drawn for a representative subject of each type, H and L, playing against the average strategies in each of the two parameterizations (averaged over all sessions). Figures 9 and 10 are contour plots reproducing the loss from not best responding, as a percentage of the maximum possible payoff. Figure 9 refers to $q = 1/2$ and Figure 10 to $q = 1/3$; in both cases the left panel refers to an H type, and the right panel to an L type. The horizontal axes in the two panels correspond to the message choices, $\widehat{\tau}_H$ or $\widehat{\tau}_L$; the vertical axis to the acceptance decisions, α or β. The shades of the different contours indicate the expected loss, from below 2.5 percent for the lightest shade, to above 25 percent for the darkest. The circles superimposed on the plots correspond to individual subject observations, with the area of the circle proportional to the number of subjects with choices at the specific point in the plot. In each panel, the red dot reports the average strategy for players of the corresponding type.

In the $q = 1/2$ sessions, there is clear asymmetry in the range of possible losses between H and L types: a maximum loss just above 10 percent of the best response payoff for H types, but higher than 20 percent for L types. For an H type, losses depend primarily on α; as $\widehat{\tau}_H$ increases, the frequency of offers of 50 declines and so does the sensitivity of expected losses to α (hence the upward sloping contours). For L types, losses can be significant if high sincerity (high $\widehat{\tau}_L$) is matched with low compliance (low β). Note that full sincerity ($\widehat{\tau}_L = 1$) and full compliance with the mediator ($\beta = 1$) are best response strategies in the data, given the behavior of others. But this is not true for H types: a sincere H type does better by rejecting 50, as the equilibrium analysis suggested. The loss however is small.

In the $q = 1/3$ sessions, there is no asymmetry in potential losses between H's and L's. H types are never offered 50 if sincere; hence the value of α makes little difference at high $\widehat{\tau}_H$. As sincerity declines, accepting 50 is increasingly costly, with potential losses reaching 15 percent at $\widehat{\tau}_H = 0$ and $\alpha = 1$. L types are only offered 30 if sincere; thus β has no impact on expected losses at low $\widehat{\tau}_L$. At higher

[35] Again conflating α_l and α_h if $q = 1/2$.

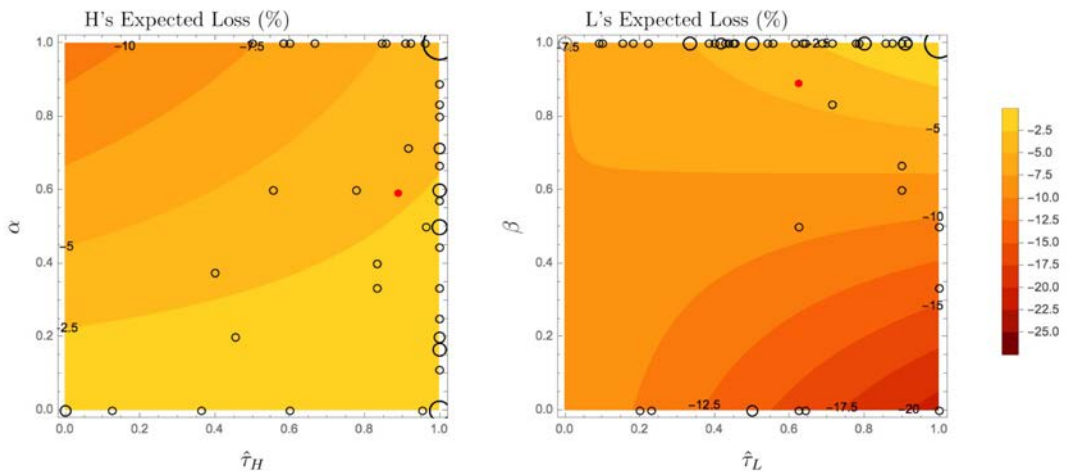

Figure 9: Losses relative to best responding; $q = 1/2$.

sincerity, however, accepting 30 becomes a preferable choice, and at $\widehat{\tau}_L = 1$ losses are monotonically declining in β. With $q = 1/3$, for both types full sincerity and compliance with the mediator's recommendations are the payoff maximizing choices in the lab. In the absence of obfuscation, as is the case for the mediator program with $q = 1/3$, lack of full sincerity by others does not affect the optimal strategies. Some robustness is built into the mediation mechanism.

The contour plots show that in both parameterizations, both types of players tend to play a pure strategy on one dimension and randomize on the other. What is interesting is that for L types behavior is quite consistent across values of q: L types in the lab predominantly accept 30 ($\beta = 1$) and randomize on the message ($\tau_L \in [0, 1]$). H types, on the other hand, change behavior with q: at $q = 1/2$, they are predominantly sincere ($\tau_H = 1$) and randomize on accepting 50 ($\alpha \in [0, 1]$); at $q = 1/3$, they randomize on the message ($\tau_H \in [0, 1]$) and predominantly reject 50 ($\alpha = 0$). The contour plots highlight in very transparent manner H types' double deviation under $q = 1/3$.

The plots also make clear that, for both values of q, the deviations from theoretical predictions we saw in the lab came at little cost. With $q = 1/2$, 93 percent of H subjects and just below two thirds (64 percent) of L subjects lost less than 5 percent from their failure to best respond to the empirical frequency of their opponents' play. With $q = 1/3$, the corresponding fractions are 92 percent for H subjects, and again 64 percent for L subjects.

The observation raises a question: are losses low because the range of possible losses is limited, or

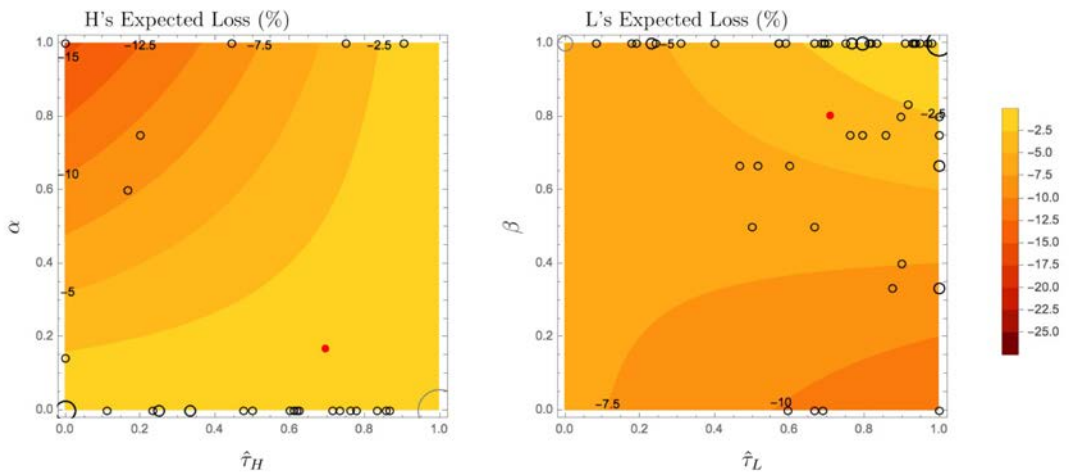

Figure 10: Losses relative to best responding; $q = 1/3$.

because subjects choose strategies that limit their losses? How badly would subjects have fared if they had acted randomly? We tested the null hypothesis of random play by simulating, for each parameterization and types, random messages and random acceptances; we then ran Kolmogorov-Smirnov tests, corrected for discreteness, comparing the distributions of random messages to the distribution of observed messages, and the distributions of random acceptance decisions to the distributions of observed acceptances.[36] All eight resulting tests strongly reject the hypothesis that subjects' choices were random ($p < 0.001$ in all cases).

Figure 11 compares CDF's of losses, in the data (in red), and under random decision-making (in grey) for each player's type and the two parameterizations.

The figure shows clearly the higher frequency of small losses in the data. With the exception of H players when $q = 1/2$, where, as shown by the contour plots, potential losses are always limited, experimental subjects are experiencing much lower losses than erratic play would induce. If subjects were playing randomly, the fractions of L players experiencing losses of not more than 5 percent would be 11 percent when $q = 1/2$ and 21 percent when $q = 1/3$, as opposed to 64 percent in the data in

[36] Ignoring silence, both messages and acceptances are binary variables–each individual observation is either 0 or 1–and because of randomness, the data show different numbers of realizations for different subjects. For each choice $c = m, a$ we thus have a sample of $n(c, t, q)$ individuals, with k_i realizations for each individual i. We construct a corresponding random data set by drawing a sample of n subjects, each with k_i realizations equally likely to be 0 or 1. We compare the empirical distribution to the random distribution via a KS test. We repeat the procedure 1,000 times. The p-value we report is the fraction of KS tests reporting a probability higher than 5 percent that the samples are drawn from the same population.

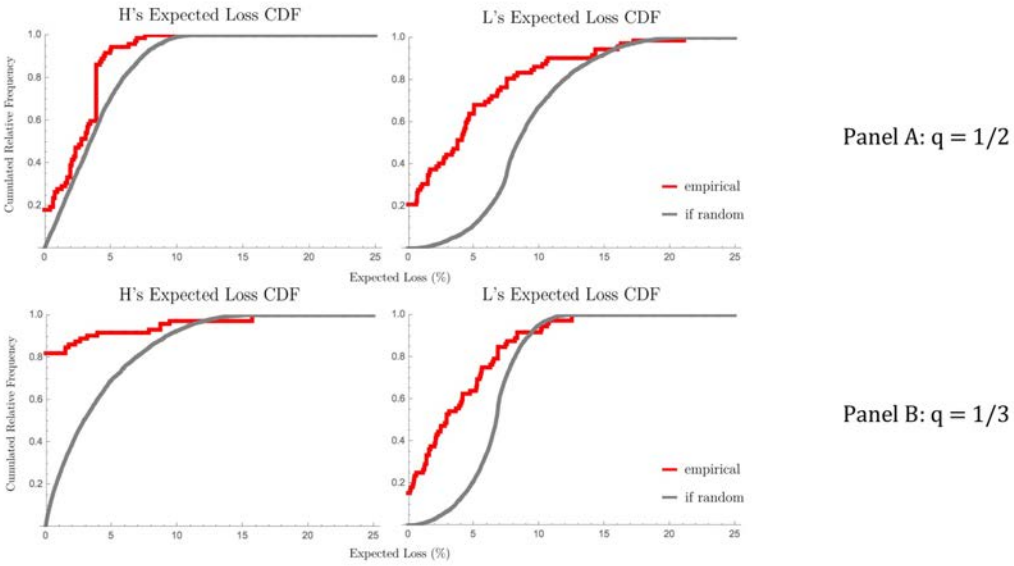

Figure 11: CDF's of losses, given observed play by others.

both cases; the equivalent numbers for H players are 70 percent with $q = 1/2$ (v/s 93 percent in the data) and 69 percent with $q = 1/3$ (v/s 92 percent in the data). Experimental subjects are playing strategies that although not best responses are not far from them, in payoff space.

6 The HM and UC treatments

The focus of the experiment was the CM treatment, but evaluating the effectiveness of the optimal mediation program requires the comparison to other treatments. Propositions HMS, 1, and 2 yield our main theoretical predictions. Beyond the broad results in the propositions, we have studied equilibria of the UC and HM games played in the lab. For both treatments, the main challenge is the large number of possible equilibria. In this section, we report briefly on some such equilibria and discuss how they fare, relative to the experimental data. All details of the derivations are in the online Appendix (Sections 9.1 and 9.3).

6.1 Human Mediation

The HM treatment was exploratory–without a known mediation program and without repeated interactions, the problem for the subjects is very difficult. It is interesting to see how subjects were

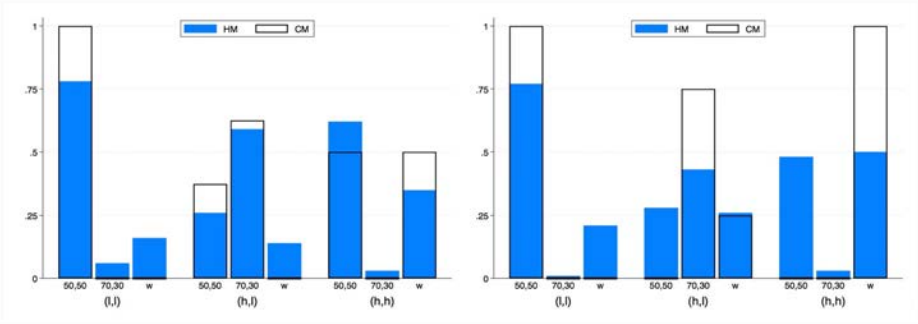

Figure 12: HM v/s CM: Recommendations.

able to use mediation under these conditions. Figures 2 and 3 as well as Tables 1 and 2 provide some answers to this question: relative to CM, sincerity is lower under HM, especially for L subjects, and so is peace, especially under $q = 1/3$.

Here we begin by comparing the mediation programs under HM and CM in Figure 12. The possible message pairs are aligned on the horizontal axis.[37] For each pair, the figure shows the frequency of different recommendations seen in the data under HM (the blue columns) and programmed into the computer mediator under CM (the black profiles).

The two programs are qualitatively similar. In particular, note HM's more frequent refusals to mediate (w) at lower q, in line with the optimal program. Refusals to mediate are not consistently higher than under CM, minimizing concerns about distortions due to risk aversion. If anything, the clearest deviation is a bias towards peace after (h, h) messages, especially under $q = 1/3$.

Table 4 reports the players' strategies in the lab in the two mediation treatments, as well as the realized peace frequencies (P). For both values of q, the players' strategies are similar, but for two systematic differences: in HM, L is less sincere ($\widehat{\tau}_L$ is smaller), and H is more willing to accept 50, especially after message l (α_l is higher).

[37] Silence in HM is rare, around 7 percent for both types under both parameterizations; in the figure we translate it into either h or l according to the prior, as happens automatically under CM.

$q = 1/2$

	$\widehat{\tau}_H$	$\widehat{\tau}_L$	α_h	α_l	β_l	β_h	P
HM	0.84	0.41	0.68	0.52	0.88	0.67	0.50
CM	0.89	0.62	0.64	0.28	0.89	–	0.55

$q = 1/3$

	$\widehat{\tau}_H$	$\widehat{\tau}_L$	α_h	α_l	β_l	β_h	P
HM	0.68	0.54	0.54	0.50	0.69	0.58	0.40
CM	0.69	0.71	–	0.17	0.80	–	0.46

Table 4. Observed players' strategies and outcomes; HM and CM.

The obvious question then is whether these regularities reflect equilibrium behavior in the HM game. The game has many equilibria; we focus here on two types of Perfect Bayesian equilibria in which dominant acceptance strategies are followed (both players accept 70, L accepts 50, and H rejects 30), and, in line with the data, $r(l,l) = (50,50)$ and H players are sincere ($\tau_H = 1$).[38] The first type of equilibria exists for both parameterizations; the second exists only for $q = 1/2$. (The precise characterization is in the online Appendix (Section 9.1)).

Equilibria 1: $q = 1/2$ and $q = 1/3$. *The mediator recommendations are:* $r(l,l) = (50,50)$, $r(h,l) = (70,30)$, $r(h,h) = w$, *and, when a recommendation is made, it is always accepted. H types are fully sincere ($\tau_H = 1$). With $q = 1/2$, L types are fully sincere as well ($\tau_L = 1$), and $P = 3/4$. With $q = 1/3$, $\tau_L = 2/3$, and $P = 56/81 \approx 0.69$.*

Hence, when $q = 1/2$, there is an equilibrium with full sincerity and full compliance. With $q = 1/3$, this cannot be sustained, but an equilibrium exists in which the mediator's recommendations, conditional on messages, are unchanged, and the players always comply, but L players are only sincere with probability 2/3.

When $q = 1/2$, there is less of a need to discipline L types' messages via refusals to mediate, and equilibria exist in which the mediator does not walk out with probability one after messages (h,h):

Equilibria 2: $q = 1/2$. *The mediator recommendations are:* $r(l,l) = (50,50)$, $r(h,l) = (70,30)$, $r(h,h) = w$ *with probability* $p_w = 0.535$ *and* $r(h,h) = (50,50)$ *with probability* $1 - p_w$*; when a*

[38] The focus on equilibria with $\tau_H = 1$ has a stronger rationale. It can be shown that any equilibrium in which $r(l,l) = (50,50)$ with probability 1, $r(h,l) = (70,30)$ with positive probability, $\tau_L > 0$, and $\tau_H > 1 - \tau_L$ (H says h more often than does L) involves full sincerity from the H type.

recommendation is made, L types always accept it; a recommendation of 50 *is accepted by H types with probability* $\alpha_h \approx 0.58$*; at the message stage:* $\tau_H = 1$ *and* $\tau_L \approx 0.565$*.* $P \approx 0.605$.[39]

Relative to observed actions in Figure 12 and Table 4, equilibria 2 for $q = 1/2$ provide plausible qualitative predictions, although HM in the lab refuses to mediate somewhat too rarely in response to (h, h), L subjects lie more than expected, and recommendations are too often rejected, inducing lower peace than predicted. Equilibria 1 find little support in the data.

6.2 Unmediated communication

The UC game is a Nash demand game with cheap talk messages. It admits a large number of equilibria, even when restricting attention to equilibria in undominated strategies. Here we discuss equilibria with properties that are desirable in a lab experiment: they are simple, in the sense that demand strategies are either conditioned on type only, or on type and a single set of messages (as opposed to both messages sent and received); they include equilibria where messages are uninformative and thus also hold in the absence of communication; and finally they yield a stark and very useful prediction: although each equilibrium set is large, the probability of peace is constant across the whole set. Proposition A2 in the online Appendix (Section 9.3) characterizes the equilibria below for arbitrary q and θ. Here we describe their main qualitative features, when specialized to the experimental parameters.

For $q = 1/3$, a particularly intuitive class of equilibria exists in which demand strategies are pure and do not depend on messages: H types always demand 70, and L types always demand 50. The equilibria impose constraints on the posterior probabilities of opponent's types, given messages, and these constraints limit the range of acceptable message strategies. Denoting by $\delta_d(T, m, m')$ the probability that type T who sent message m and received message m' demands d:

Equilibria 1: $q = 1/3$. *At the demand stage:* $\delta_{70}(H, m, m') = 1$, $\delta_{50}(L, m, m') = 1$ *for all* m, m'*. At the message stage, for any* $(\tau_L + \sigma_L) \in (0, 1)$, $\tau_H \in [\max(0, 1-\sigma_H-(4/3)\tau_L), \min(1, (4/3)(1-\tau_L-\sigma_L))]$, $\sigma_H \leq (4/3)\sigma_L$*.* $P = (1-q)^2 = 4/9 \approx 0.444$.

Note that the equilibrium set includes fully uninformative messages ($\tau_H = 1 - \tau_L - \sigma_L$, $\sigma_L = \sigma_H$), and the demand strategies remain equilibrium strategies of the game without communication. In these equilibria communication, even when informative, does not affect peace: the peace probability

[39] As shown in the online Appendix, the strategies described can be supported by different off-equilibrium acceptance strategies. Hence the plural term "equilibria".

is constant over the full range of acceptable messages.

	$q = 1/3$							
	P	τ_H	σ_H	$\delta_{70}(H)$	$\delta_{70}(L,l)$	$\delta_{70}(L,h)$	$\delta_{70}(L,s)$	$\delta_{50}(L)$
data	0.49	0.74	0.10	0.71	0.11	0.21	0.13	0.72
equil1	0.44	$[0.5, 0.74]$	≤ 0.19	1	0	0	0	1
equil2	0.35	$[0.19, 0.80]$	$[0.05, 0.39]$	1	0.91	0.57	0.85	0

	$q = 1/2$							
	P	τ_H	σ_H	$\delta_{70}(H)$	$\delta_{70}(L,l)$	$\delta_{70}(L,h)$	$\delta_{70}(L,s)$	$\delta_{50}(L)$
data	0.57	0.81	0.12	0.52	0.02	0.15	0.14	0.73
equil2	0.59	$[0.47, 0.80]$	$[0.02, 0.13]$	1	0.95	0	0.07	0

Table 5. Unmediated Communication: Data and Equilibria.

As shown in Table 5, this class of equilibria explains the data reasonably well, whether in terms of message and demand strategies, or peace. Such equilibria however do not exist for $q = 1/2$. We need to allow L's demand strategies to be mixed:

Equilibria 2: $q = 1/2$ and $q = 1/3$. *At the demand stage:* $\delta_{70}(H, m, m') = 1$ *for all* m, m'; $\delta_{70}(L, m, m') = 2\left[1 - \frac{3}{7}\left(\frac{1}{1-\pi_m}\right)\right] = 1 - \delta_{30}(L, l, m')$ *for all* m', *where* π_m *is the posterior probability a player is* H *given message* m.[40] *At the message stage, for any* $(\tau_L + \sigma_L) \in (0, 1)$, $\tau_H \in [\underline{\tau_H}(\tau_L, \sigma_H, \sigma_L), \overline{\tau_H}(\tau_L, \sigma_H, \sigma_L)]$ *and* $\sigma_H \in [\underline{\sigma_H}(\sigma_L), \overline{\sigma_H}(\sigma_L)]$. $P = 0.586$ if $q = 1/2$, and 0.345 if $q = 1/3$.

The boundaries on the message strategies are not very illuminating, and we leave them to the online Appendix. For both values of q, they include fully uninformative messages, implying that demand strategies remain equilibria in the game without communication. In fact, for $q = 1/2$, these are the only equilibrium demand strategies that exist without communication. Again, even when messages are informative, the ex ante peace probability remains constant: the mixing probabilities at the demand stage effectively nullify the information provided by the message.

[40] Hence for example

$$\pi_h = \frac{q\tau_H}{q\tau_H + (1-q)(1-\tau_L-\sigma_L)}.$$

The probability of peace is lower with $q = 1/3$ than in the case of $q = 1/2$ under both sets of equilibria: the smaller likelihood that the opponent is H makes L types more aggressive and results in more frequent conflict. Because of the large possible range of equilibrium messages, and hence of mixing probabilities over demands, lower peace under $q = 1/3$ is the only treatment effect that can be claimed with some confidence. It is indeed observed in the data (see Table 5).[41]

The predictions of the second class of equilibria are reported in Table 5 as well, with demand strategies anchored by the observed point values for τ_L, τ_H, σ_L, and σ_H. The second set of equilibria cannot explain demand strategies, especially for L subjects.

We conclude this section by representing graphically our data from the three treatments in the 3D figure we have used repeatedly, reporting sincerity (adjusted for silence) on the horizontal axes, and peace on the vertical axis. Each sphere in Figure 13 corresponds to an experimental session; red spheres correspond to CM, blue to HM and yellow to UC. The figure shows clearly the increase in L's sincerity in moving from UC to HM to CM, the constancy of the peace frequency across the treatments, and the higher variance in peace across sessions for HM, probably mirroring the high complexity of the treatment.

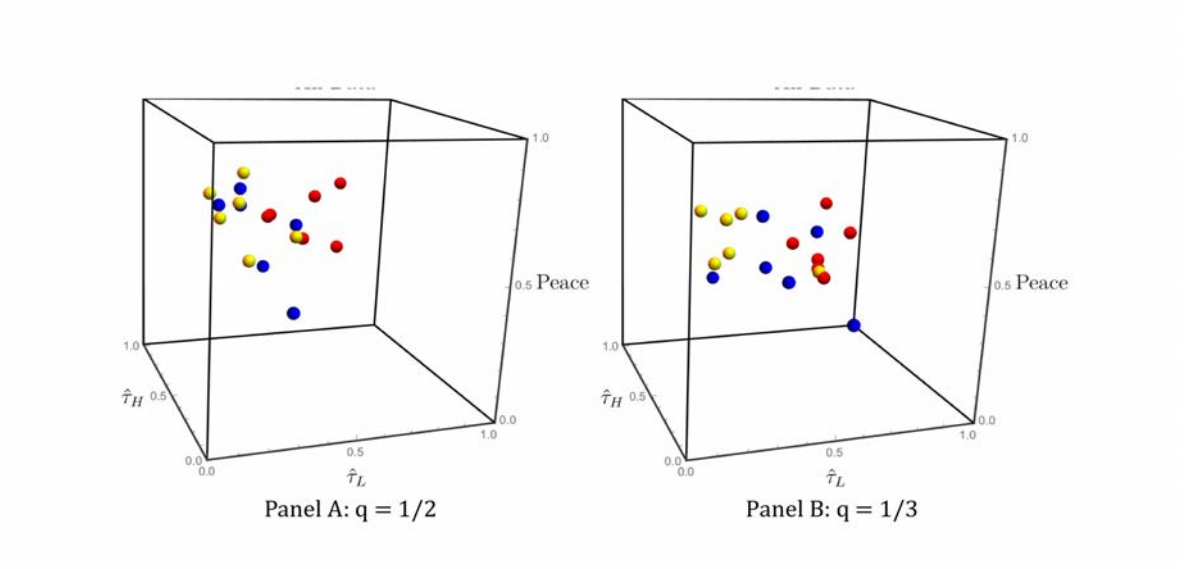

Figure 13: All treatments: UC, HM, and CM.

[41] In addition, the hypothesis of equal probability of peace in the absence of communication is supported by the data from the initial experimental rounds with no communication. See the discussion of the NC data in online Appendix (Section 9.4.4).

7 Conclusions

The experiment tests the potential of a sophisticated mediation mechanism in reducing conflict between two parties whose strength is privately known. The mediator has no superior information, no independent resources, and no enforcement power. Yet, theory predicts that mediation can lead to a strictly lower frequency of conflict than if the two parties must communicate directly. We implement the optimal mediation mechanism in the lab and find that, in line with the theory, participants reveal their strength more sincerely than when communicating directly; however, contrary to the theory, the frequency of conflict is not lower.

Multiple equilibria are partly responsible for the result, but so are two other less expected factors. First, theory suggest that the superior outcome is reached when the mediator's recommendation leaves the parties unsure of the opponent's strength. We find that it is exactly in this case that the optimal equilibrium is especially fragile: in the neighborhood of the highest-peace equilibrium, the locus of equilibria is discontinuous in outcomes, and any positive probability of lying by the opponent, no matter how small, comes with a discontinuous upward jump in the frequency of conflict. The jump is due to non-compliance with some of the mediator's recommendations. In the lab, sincerity is not perfect and, as predicted, neither is compliance.

Second, whether or not the optimal mediation program reveals fully the parties' strength depends on parameter values. We test mediation under both parameterizations and find that in both the actions participants take in the lab have some noise. The actions we observe are not erratic and deviations from best responses cause participants only small payoff losses. Yet, this "good enough" play has significant repercussions on the frequency of conflict.

We come away from the experiment with two main lessons. First, under the optimal mediation mechanism the incentive constraints apply with no slack, and the equilibrium is not strict: any slack in the constraints is suboptimal because it can be reduced while increasing the probability of peace. In the lab, however, modifying the mediation program so that the best equilibrium is strict could improve its performance. It would be interesting to think more about how best to do so, and to test the modified mechanism. Second, in the lab and in the world, some noise in behavior is to be expected. Because the sources of noise can be quite diverse, the question of the robustness of a mechanism to behavioral noise is inherently experimental. In our experiment, as we noted, the noise we see has small effects on individual payoffs, relative to best responding, and we cannot attribute the

results to participants' confusion. Yet, the mediation mechanism requires complex strategic thinking. It would be interesting to think of designing mediation mechanisms that are strategically simple (as discussed in Li, 2017 or Börgers and Li, 2019, for example) or rely on boundedly rational thinking (as in Kneeland, 2020, or de Clippel et al, 2019, for example) and to test their performance relative to the theoretically optimal mechanism with fully rational agents.

References

Aghion, P., E. Fehr, R. Holden and T. Wilkening, 2018, "The Role of Bounded Rationality and Imperfect Information in Subgame Perfect Implementation—An Empirical Investigation", *Journal of the European Economic Association*, 16, 232-274.

Aristidou, A., G. Coricelli, and A. Vostroknutov, 2019, "Incentives or Persuasion? An Experimental Investigation," Research Memorandum 012, Maastricht University, Graduate School of Business and Economics (GSBE).

Attiyeh, G., R. Franciosi and M. Isaac, 2000, "Experiments with the Pivot Process for Providing Public Goods", *Public Choice*,102, 93–112.

Au, P. H. and K. K. Li, 2018, "Bayesian Persuasion and Reciprocity: Theory and Experiment", https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3191203

Banks, J., M. Olson, D. Porter, S. Rassenti and V. Smith, 2003, "Theory, Experiments and FCC Spectrum Auctions", *Journal of Economic Behavior and Organization*, 51, 303–350.

Beardsley, K., 2011, *The Mediation Dilemma*, Ithaca, NY: Cornell University Press.

Bester, H. and R. Strausz, 2000, "Imperfect Commitment and the Revelation Principle: the Multi-Agent Case", *Economics Letters*, 69, 165-171.

Bester, H. and R. Strausz, 2001, "Contracting with Imperfect Commitment and the Revelation Principle: the Single Agent Case", *Econometrica*, 69, 1077-1098.

Blume, A., E. K. Lai and W. Lim, 2019, "Mediated Talk: An Experiment",

http://wooyoung.people.ust.hk/Mediated%20Talk%20Experiment-February-11-2019.pdf.

Börgers, T. and J. Li, 2019, "Strategically Simple Mechanisms", *Econometrica*, 87, 2003–2035.

Brown, J. and I. Ayres, 1994, "Economic Rationales for Mediation", *Virginia Law Review*, 80, 323-402.

Brunner, C., J. Goeree, C. Holt and J. Ledyard, 2010, "An Experimental Test of Flexible Combinatorial Spectrum Auction Formats", *AEJ: Microeconomics*, 2, 39 – 57.

Cason, T., T. Saijo, T. Sjöström and T. Yamato, 2006, "Secure Implementation Experiments: Do Strategy-Proof Mechanisms Really Work?", *Games and Economic Behavior*, 57, 206-235.

Chen, Y., 2008, "Incentive-Compatible Mechanisms for Pure Public Goods" A Survey of Experimental Research", in C. Plott and V. Smith (eds.), *The Handbook of Experimental Economics Results*, 625-643, New York, NY: North-Holland.

Chen, Y and C. Plott, 1996, "The Groves–Ledyard Mechanism: An Experimental Study of Institutional Design", *Journal of Public Economics*, 59, 335–364.

Chen, Y. and T. Sonmez, 2006, "School Choice: An Experimental Study", *Journal of Economic Theory*, 127, 202-231.

Cornich, C., 2019, "Industry of peacemakers capitalizes on global conflict", *The Financial Times*, October, 22.

de Clippel, G., R, Saran and R. Serrano, 2019, "Level-k Mechanism Design," *The Review of Economic Studies*, 86, 1207–1227.

Fanning, J., 2019, "Mediation in Reputational Bargaining",

https://sites.google.com/a/brown.edu/jfanning.

Fey, M. and K. Ramsay, 2010, "When is Shuttle Diplomacy Worth the Commute? Information Sharing through Mediation", *World Politics*, 62, 529-60.

Fischbacher, U., 2007, "z-Tree: Zurich Toolbox for Ready-made Economic Experiments", *Experimental Economics*, 10, 171–178.

Forges, F. ,1986, "An Approach to Communication Equilibria", *Econometrica*, 54, 1375–1385.

Forges, F., 1990, "Equilibria with Communication in a Job Market Example", *Quarterly Journal of Economics*, 105, 375-398.

Frechette, G., A. Lizzeri, and J. Perego, 2019, "Rules and Commitment in Communication", CEPR D.P. No.14085.

Goltsman, M., J. Hőrner, G. Pavlov, and F. Squintani, 2009, "Mediation, Arbitration and Negotiation", *Journal of Economic Theory* 144, 1397–1420.

Greiner, B., 2015, "Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE". *Journal of the Economic Science Association*,1, 114–125.

Hörner, J., M. Morelli and F. Squintani, 2015, "Mediation and Peace", *Review of Economic Studies* 82, 1483–1501.

Kneeland, T., 2020, "Mechanism Design with Level-k Types: Theory and an Application to Bilateral Trade",

http://www.tkneeland.com/uploads/9/5/4/8/95483354/levelk_mechanismdesign_04.06.2020.pdf.

Krishna, V, 2007, "Communication in games of incomplete information: Two players", *Journal of Economic Theory*, 132, 584-592.

Kydd, A., 2003, "Which Side are You on? Mediation as Cheap Talk", *American Journal of*

Political Science, 47, 596–611.

Kydd, A., 2006, "When Can Mediators Build Trust?", *American Political Science Review*, 100, 449-462.

Li, S., 2017, "Obviously Strategy-Proof Mechanisms", *American Economic Review*, 107, 3257-87.

Meirowitz, A., M. Morelli, K. Ramsay and F. Squintani, 2019, "Dispute Resolution Institutions and Strategic Militarization", *Journal of Political Economy*, 127, 378-418.

Myerson, R., 1996, *Game Theory: Analysis of Conflict*, Cambridge, Ma and London, UK: Harvard University Press.

Nguyen, Q., 2017, "Bayesian Persuasion: Evidence from the Laboratory," Working Paper.

Palfrey, T., 1990, "Implementation in Bayesian Equilibrium: The Multiple Equilibrium Problem in Mechanism Design", Social Science W.P. No.760, Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA.

Roth, A., 2016, "Experiments in Market Design" in J. Kagel and A. Roth (eds.), *The Handbook of Experimental Economics*, vol. 2, 290–346, Princeton, NJ: Princeton University.

Wilkenfeld, J., K. Young, V. Asal, and D. Quinn, 2003, "Mediating International Crises: Cross-National and Experimental Perspectives", *Journal of Conflict Resolution*, 47, 279–301.

8 Appendix

8.1 Mediation without Commitment: Proposition 2

Proposition 2. *Assume $q < (2\theta - 1)/\theta$ and $q \neq 2\theta - 1$. If the mediator cannot commit to refuse mediation, any truthful equilibrium involves a probability of peace that is strictly lower than can be achieved by a mediator with commitment power.*

In line with the experimental design described later, we call the mediator without commitment the Human Mediator (HM), and we focus on symmetric Perfect Bayesian Equilibria (PBE) in undominated strategies (both players accept θ, L accepts 1, and H rejects $(1-\theta)$) which we refer to as "equilibria". We begin by establishing the following lemma, which characterizes all truthful equilibria. For the result, we allow for general HM payoffs: she receives 1 if an offer is accepted, 0 if an offer is rejected, and $W \in [0, 1)$ for walking out.

Lemma A1. *Assume $q < (2\theta - 1)/\theta$. The following characterizes all truthful equilibria (on-path strategies and outcomes), which fall into one of two families.*

Equilibria 1: $q \geq 2\theta - 1$ and $W \in [0, 1)$. *(i) Following messages (l, l), HM mixes arbitrarily between offering $(1/2, 1/2)$ and $(\theta, 1-\theta)$ (randomizing which player is offered θ). The offer is always accepted. (ii) Following (h, l), HM offers $(\theta, 1-\theta)$, which is always accepted. (iii) Following (h, h), if $W > 0$, HM walks out; if $W = 0$, HM mixes arbitrarily between walking out or making any offer, but the offer is always rejected.*

Equilibria 2: $W = 0$. *(i) Following messages (l, l), HM offers $(1/2, 1/2)$. (ii) Following (h, l), HM mixes arbitrarily between walking out and offering $(1/2, 1/2)$, which is always rejected. (iii) Following (h, h), HM mixes arbitrarily between walking out or making any offer, but the offer is always rejected.*

Proof. Suppose both types of player are sincere. In any PBE in undominated strategies, L accepts $1/2$. Thus, following (l, l), HM can maximize payoffs by offering $(1/2, 1/2)$ which will be accepted with probability 1 (w.p. 1); hence HM will never walk out. Following (l, l), HM can offer $(\theta, 1-\theta)$ with positive probability only if it is accepted w.p. 1.

Following (h, l), first suppose that HM offers $(\theta, 1-\theta)$ w.p. > 0. In this case, it will be accepted w.p. 1. To see this, it must be that either $(\theta, 1-\theta)$ is offered w.p. 0 or w.p. > 0 following (l, l). In the former case, $(\theta, 1-\theta)$ offered to (h, l) reveals to L that the opponent is H, and so she accepts. In the latter case, L must accept since she cannot distinguish between the information sets following (h, l) and (l, l), and, as argued, $(\theta, 1-\theta)$ must be accepted in equilibrium if offered following (l, l). Hence,

HM does not walk out following (h,l) w.p. >0. Can it be that HM mixes between $(\theta,1-\theta)$ and $(1/2,1/2)$ (both with positive probability) following (h,l)? Since $(\theta,1-\theta)$ would be accepted w.p. 1, $(1/2,1/2)$ can only be offered with positive probability if it is accepted w.p. 1. But then, following (h,h), HM would offer $(1/2,1/2)$ w.p. 1 since it would be accepted w.p. 1 (and $(\theta,1-\theta)$ would be rejected w.p. 1 since H's always reject $1-\theta$). But then, it would not be an equilibrium for L to be sincere (by saying h, L would receive a strictly higher payoff if the other player is H). Thus, it must be that, if (h,l) is followed by $(\theta,1-\theta)$ with positive probability, then the probability must equal 1. In other words, (h,l) must be followed by $(\theta,1-\theta)$ w.p. 1 or w.p. 0.

We now search for equilibria in each of 4 cases. We characterize equilibria without considering the option of messaging s, but since being sincere is part of these equilibria, if s would be interpreted as h with any probability q_s and l with probability $1-q_s$, deviation to s cannot be advantageous for either player.

Case 1: $W>0$ and (h,l) is offered $(\theta,1-\theta)$ w.p. 1.

Following (h,h), an offer of $(\theta,1-\theta)$ will be rejected (as H will always reject $1-\theta$). Can HM offer $(1/2,1/2)$ with positive probability following (h,h)? If sincere H players expect others sincere H players to accept $(1/2,1/2)$ with positive probability, then it is uniquely optimal for them to accept. But then offering $(1/2,1/2)$ is uniquely optimal for HM. But this cannot be part of a sincere equilibrium because then L would strictly prefer to lie. Hence, HM must choose w following (h,h).

Recall that in this candidate equilibrium, HM offers $(1/2,1/2)$ with some probability $p50_{ll}\in[0,1]$ and $(\theta,1-\theta)$ with probability $1-p50_{ll}$. Note that if such an equilibrium exists for an arbitrary $p50_{ll}\in[0,1]$, then the equilibrium exists for all $p50_{ll}\in[0,1]$: following (l,l), both $(1/2,1/2)$ and $(\theta,1-\theta)$ are always accepted (since (h,l) is offered $(\theta,1-\theta)$ w.p. 1, it is optimal for L to accept $(1-\theta)$), and L's expected payoff from messaging l against a sincere L opponent is $p50_{ll}(1/2)+(1-p50_{ll})[\theta/2+(1-\theta)/2]=1/2$ for all $p50_{ll}$.

Note also, by construction of the candidate equilibrium strategies, HM is optimizing given others' behavior and the players's acceptance strategies are optimal. It remains to check that being sincere is optimal for both player types.

After messaging h, H is offered θ against L, and accepts, and is brought to war against H, for a total expected payoff of $(1-q)\theta+q\theta/2$. If H messages l, H is offered $(1-\theta)$ when opposite a sincere H, and rejects, and either $1/2$ or $(1-\theta)$ or θ when opposite a sincere L, and rejects all but θ, for the same total expected payoff of $(1-q)\theta+q\theta/2$. It remains to check that sincerity is a best response

for the L-type. Under sincerity, L's expected payoff is $q(1-\theta)+(1-q)(1/2)$. If L sends message h, L's expected payoff is $q(0)+(1-q)\theta \leq q(1-\theta)+(1-q)(1/2) \iff q \geq 2\theta-1$. Thus, we have characterized equilibria for this case, which exist if and only if $q \geq 2\theta-1$. These are summarized in "Equilibria 1".

Case 2: $W>0$ and (h,l) is offered $(\theta,1-\theta)$ w.p. 0.

Can it be that (h,l) is followed by $(1/2,1/2)$ w.p. 1? If (h,l) is followed by $(1/2,1/2)$ w.p. 1, it will be rejected w.p. 1 by the H type. To see this note that an H will accept $1/2$ only if:

$$\begin{gathered}
Pr(H_j|(1/2,1/2))Pr(H_j \text{ accepts } 1/2)(1/2-\theta/2)+ \\
Pr(L_j|(1/2,1/2))(1/2-\theta) \geq 0 \iff \\
Pr(H_j|(1/2,1/2))Pr(H_j \text{ accepts } 1/2)(1/2-\theta/2) \geq Pr(L_j|(1/2,1/2))(\theta-1/2) \iff \\
Pr(1/2,1/2|hh)qPr(H_j \text{ accepts } 1/2)(1-\theta) \geq Pr(1/2,1/2|hl)(1-q)(2\theta-1).
\end{gathered}$$

If $Pr(1/2,1/2|hl)=1$, the condition becomes:

$$Pr(1/2,1/2|hh)Pr(H_j \text{ accepts } 1/2) \geq \left(\frac{1-q}{q}\right)\left(\frac{2\theta-1}{1-\theta}\right).$$

But:

$$\left(\frac{1-q}{q}\right)\left(\frac{2\theta-1}{1-\theta}\right)>1 \iff q<\frac{2\theta-1}{\theta},$$

which is satisfied in the model. Hence if (h,l) is followed by $(1/2,1/2)$ w.p. 1, $(1/2,1/2)$ is always rejected by the H type. Hence, HM will prefer to walk out and receive $W>0$.

Can it be that, following (h,l), HM mixes between $(1/2,1/2)$ and w (both with positive probability)? The answer is no. To see this, it must be that HM is indifferent between the two offers, which requires

$$Pr(H_i \text{ accepts } 1/2)=W \in (0,1),$$

and thus it must be that H is indifferent between accepting and rejecting $\frac{1}{2}$. Inspecting H's indifference condition from above and noting that offering $(\theta,1-\theta)$ following (h,h) will be rejected w.p. 1 (as H will always reject $1-\theta$), to keep H indifferent requires that HM mixes between $(1/2,1/2)$ and w following both (h,l) and (h,h). However, this cannot be optimal for HM: if she is indifferent between $(1/2,1/2)$ and w following (h,l), she strictly prefers w following (h,h). Hence, there can be

no equilibria for this case.

Case 3: $W = 0$ and (h, l) is offered $(\theta, 1-\theta)$ w.p. 1.

This is exactly the same as Case 1 except now, following (h, h), since HM is indifferent between walking out and having an offer rejected, she may choose any mixture of w, an offer $(\theta, 1-\theta)$ which will be rejected, or an offer $(\frac{1}{2}, \frac{1}{2})$ which must be rejected w.p. 1 as part of the equilibrium. These equilibria are summarized in "Equilibria 1".

Case 4: $W = 0$ and (h, l) is offered $(\theta, 1-\theta)$ w.p. 0.

Can it be that (h, l) is followed by an arbitrary mixture of $(1/2, 1/2)$ and w? In order for HM to be willing to mix, H must reject $\frac{1}{2}$ w.p. 1, which is optimal for H if other H's do the same. Hence, HM is willing to mix over of $(1/2, 1/2)$ and w following (h, l) if deviating and offering $(\theta, 1-\theta)$ is rejected w.p. 1. This can be part of the equilibrium if (l, l) is offered $(1/2, 1/2)$ w.p. 1: if (l, l) is offered $(1/2, 1/2)$ w.p. 1, L being offered $1-\theta$ is off-path and so her accepting can be supported by the belief that her opponent is H. Following (h, h), it must be that $(1/2, 1/2)$ is rejected w.p. 1 since H can not distinguish information sets following (h, h) and (h, l). Since $(\theta, 1-\theta)$ would also be rejected (as H will always reject $1-\theta$), HM may mix arbitrarily between any offer or walking out following (h, h) which leads to war.

By construction, HM is optimizing given others' behavior, the players's acceptance strategies are optimal, and it is easy to check that being sincere is optimal for both player types. Thus, we have characterized equilibria for this case. These are summarized in "Equilibria 2". □

Proof of Proposition 2. Equilibria 1, which only exist for $q \geq (2\theta - 1)$, involve peace $P_1 = 1 - q^2$. Equilibria 2, which only exist for $W = 0$, involve peace $P_2 = (1-q)^2 < P_1$. In the optimal equilibrium under mediation with commitment, the probability of peace is $\theta(1-q)^2/(\theta - q)$ (from HMS's Lemma 3), which is always greater than P_2. What about P_1? $P_1 = 1 - q^2 < \theta(1-q)^2/(\theta - q) \iff (2\theta - 1)/\theta > q > (2\theta - 1)$. Hence, the no-commitment level of peace achieves the commitment level of peace if and only if $q = (2\theta - 1)$. □

8.2 Multiple Equilibria under CM (at the experimental parameter values)

We characterize players' equilibrium strategies keeping fixed the mediator's mechanism as programmed under CM. We consider the multi-agent representation of the extensive form game and concentrate on equilibria that satisfy two requirements. First, to avoid indeterminacies in Bayesian updating that are not relevant to explaining our experimental data, we focus on equilibria where the probability of

observing either message, l or h, is always positive, if possibly arbitrarily small (that is, we rule out the corners $(\tau_L = 0, \tau_H = 1)$ and $(\tau_L = 1, \tau_H = 0)$). [42]. Second, we select 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. Denoting by Tm a player of type T who sent message m, what remains to be determined are the acceptance strategies of Hh and Hl players offered 50, and of Ll players offered 30, as well as the first stage message strategies for both types. We simplify notation by denoting by α_m the probability of an Hm type accepting 50, and by β the probability of an Ll type accepting 30. As before, we denote by τ_T the probability that type $T \in \{H, L\}$ is truthful and by $\widehat{\tau}_T$ the probability that a message sent by type T is read as T by the computer mediator when silence is accounted for.

Because the mediation program does not include obfuscation under $q = 1/3$, that case is simpler and it is helpful to analyze it first.

8.2.1 $q = 1/3$

We begin by ignoring the option of silent messages; at the end of the subsection we show how the results generalize when silent messages are included. Consider first the acceptance decisions. When $q = 1/3$, a player announcing h faces either $r = w$, if the mediator refuses to mediate, or $r = 70$, which the player always accepts. Hence non-trivial acceptance decisions only concern Hl offered 50 and Ll offered 30. In both cases accepting is optimal if the opponent is an H, but rejecting is optimal if the opponent is L.

The mediation program has no obfuscation: given the mediator's recommendation, each player knows the message sent by the opponent. (i) Consider first an Ll offered 30 (who thus knows that the opponent, j, sent message h). The player's own acceptance strategy is relevant only if the opponent accepts. But the opponent is offered 70 and all accept 70; hence conditioning on the opponent's acceptance yields no additional information on the opponent's type. The posterior probability of the opponent's type is straightforward:[43]

$$\Pr(j \ is \ L|h_j) = \frac{2(1-\tau_L)}{2(1-\tau_L)+\tau_H} \tag{3}$$

[42] When allowing for silence, an equivalent restriction is to select equilibria with an arbitrarily small but positive probability of silence.

[43] The restriction to equilibria with positive probability of observing either message rules out $\tau_L = 1$ and $\tau_H = 0$ (all types always say l), thus guaranteeing that (3) is well-defined. A similar observation applies to other posterior probabilities below, and is not repeated.

In any equilibrium in which all accept 70, Ll player i will accept 30 with positive probability only if $EU_{Ll}(accept\ 30) \geq EU_{Ll}(reject\ 30)$ where:

$$EU_{Ll}(accept\ 30) = 30$$
$$EU_{Ll}(reject\ 30) = 35\Pr(j = L|h_j)$$

Substituting (3):

$$3\tau_H = 1 - \tau_L \Longrightarrow \beta \in [0,1] \text{ and} \tag{4}$$
$$3\tau_H < 1 - \tau_L \Longrightarrow \beta = 0,\ 3\tau_H > 1 - \tau_L \Longrightarrow \beta = 1$$

Note that since we are considering the acceptance decision for a player of type L who sent message l, in equilibrium the condition is only relevant when $\tau_L > 0$. If $\tau_L = 0$, the condition anchors off-equilibrium behavior.

(ii) Consider now an Hl who is offered 50 (and thus knows that the opponent, j, sent message l). L players always accept 50, but H players may not. Thus conditioning on j's acceptance can yield relevant information. Consider a candidate equilibrium where the Hl player expects other Hl players to accept 50 with some probability $\alpha_l \in [0,1]$. It is immediate that:

$$EU_{Hl}(accept\ 50) - EU_{Hl}(reject\ 50) =$$
$$= 15\Pr(j\ accepts\ and\ is\ H|50, l_j) - 20\Pr(j\ accepts\ and\ is\ L|50, l_j).$$

where:

$$\Pr(j\ accepts\ and\ is\ H|50, l_j) = \frac{\alpha_l(1-\tau_H)}{(1-\tau_H)+2\tau_L}, \tag{5}$$
$$\Pr(j\ accepts\ and\ is\ L|50, l_j) = \Pr(j\ is\ L|\ l_j) = \frac{2\tau_L}{2\tau_L + (1-\tau_H).}$$

Hence:

$$15\alpha_l(1-\tau_H) = 20(2\tau_L) \Longrightarrow \alpha_l \in [0,1] \text{ and} \tag{6}$$
$$15\alpha_l(1-\tau_H) > 20(2\tau_L) \Longrightarrow \alpha_l = 1,\ 15\alpha_l(1-\tau_H) < 20(2\tau_L) \Longrightarrow \alpha_l = 0.$$

As above, since we are considering the acceptance decision for a player of type H who sent message l, in equilibrium the condition is only relevant for $\tau_H < 1$. If $\tau_H = 1$, the condition anchors off-equilibrium behavior: an H type who deviated and sent message l, would anticipate that when the mediator's recommendation of $(50, 50)$ is received, his future self would know, given $\tau_H = 1$, that j is an L, and thus would reject the recommendation. Note also that $\alpha_l = 0$ is self-enforcing: if all Hl types who sent message l reject the equal split, then the only type who would accept is an L; but then Hl prefers to reject.

We can now move back to the message stage: $\tau_H = 1$ if $EU_H(h) > EU_H(l)$, and $\tau_H \in [0, 1]$ if $EU_H(h) = EU_H(l)$ (and similarly, $\tau_L = 1$ if $EU_L(l) > EU_L(h)$, and $\tau_L \in [0, 1]$ if $EU_L(l) = EU_L(h)$). Recalling that β is the probability that Ll accepts 30 and $q = 1/3$, the relevant expected utility equations are:

$$EU_H(h) = (1/3)\left[\tau_H 35 + (1-\tau_H)(35/4 + 35(3/4))\right] + (2/3)\left[\tau_L 70 + (1-\tau_L)70\right] = (1/3)35 + (2/3)70 \tag{7}$$

$$EU_H(l) = (1/3)\left[\tau_H 35 + (1-\tau_H)(\alpha_l^2 50 + (1-\alpha_l^2)35)\right] + (2/3)\left[\tau_L(\alpha_l 50 + (1-\alpha_l)70) + (1-\tau_L)70\right]$$

$$EU_L(l) = (1/3)[\tau_H(3/4)\beta 30 + (1-\tau_H)\alpha_l 50] + (2/3)[\tau_L 50 + (1-\tau_L)(35/4 + (3/4)(30\beta + 35(1-\beta)))]$$

$$EU_L(h) = (1/3)[\tau_H(0) + (1-\tau_H)(0)] + (2/3)[\tau_L(35/4 + (3/4)(\beta 70 + (1-\beta)35)) + (1-\tau_L)35]$$

Four conditions, ((3), (5), and the relevant expected utility comparisons), together with the constraints $\alpha \in [0, 1]$, $\beta \in [0, 1]$, $\tau_H \in [0, 1]$, $\tau_L \in [0, 1]$, and the no-indeterminacy conditions ($\tau_L = 0 \Longrightarrow \tau_H \neq 1$) and ($\tau_L = 1 \Longrightarrow \tau_H \neq 0$) determine the equilibrium values of α, β, τ_H and τ_L. The derivation is straightforward, although considering all different possible cases is cumbersome. Solving a specific case helps to build intuition. Suppose $\alpha_l = 0$ and $\beta = 1$. Then: (i) $EU_H(h) = EU_H(l)$ for all τ_H, τ_L–an H type is indifferent over any message; (ii) $\tau_L = 1$ if $\tau_H > 1/3 + (2/3)\tau_L$ and $\tau_L \in [0, 1]$ if $\tau_H = 1/3 + (2/3)\tau_L$; (iii) for any $\tau_L \in [0, 1]$ and $\tau_H = 1/3 + (2/3)\tau_L$, $\alpha_l = 0$ and $\beta = 1$ satisfy (6) and (4). Thus indeed there exist a continuum of equilibria such that message strategies are: $\tau_L \in [0, 1]$, $\tau_H = 1/3 + (2/3)\tau_L$; and acceptance strategies are: H types only accept 70, L types always accept all offers.

The full set of equilibria is as follows:

$$
\begin{aligned}
&(i)\ \beta = 1, \tau_L = 1, \tau_H = 1;\\
&(ii)\ \alpha_l = 0, \beta = 1, \tau_L \in (0,1), \tau_H = 1/3 + (2/3)\tau_L;\\
&(iii)\ \alpha_l = 0,\ \beta = 4/7, \tau_L \in (0,1), \tau_H = 1/3 - \tau_L/3;\\
&(iv)\ \alpha_l = 0, \tau_L = 0, \tau_H \leq 1/3.
\end{aligned}
$$

Equilibrium (i) is the HMS equilibrium.

Silence The equilibria above ignored the option of Silence. We show here that when we account for Silence all results above apply with a simple transformation of variables.

With $q = 1/3$, the mediator's mechanism has no obfuscation and thus if a recommendation is made, it reveals to each player how the mediator has read the two messages. Call $\widehat{m}$ a message read as m by the computer mediator, and recall that under treatment CM, the rule according to which silent messages are read by the computer is specified. Consider for example the problem of player i who sent message l_i, received recommendation $(30, 70)$, and wants to evaluate the probability that opponent is H. From the recommendation, player i knows that the opponent's message was $\widehat{h}$, i.e. was read as h by the computer. Then, as usual denoting by σ_T the probability that type T sends a silent message:

$$
\begin{aligned}
\Pr(j\ is\ L|\widehat{h}_j) &= \frac{\Pr(\widehat{h}_j|j\ is\ L)\Pr(L)}{\Pr(\widehat{h}_j|j\ is\ L)\Pr(L) + \Pr(\widehat{h}_j|j\ is\ H)\Pr(H)}\\
&= \frac{[1-\tau_L-\sigma_L+(1/3)\sigma_L](2/3)}{[1-\tau_L=\sigma_L+(1/3)\sigma_L](2/3)+[\tau_H+(1/3)\sigma_H](1/3)}\\
&= \frac{2(1-\widehat{\tau}_L)}{2(1-\widehat{\tau}_L)+\widehat{\tau}_H}
\end{aligned}
$$

With a change in variable, the formula is identical to (3). The conclusion extends to all results in the previous section, reinterpreted by substituting $\widehat{\tau}_H$ and $\widehat{\tau}_L$ for τ_H and τ_L. Summarizing, the computer can read the subject's true type with probability 1 only if $\sigma_T = 0$; otherwise, in equilibrium τ_T and σ_T are jointly determined. Using $\widehat{\tau}_H$ and $\widehat{\tau}_L$ in updating the opponent's expected type, given the recommendation, acceptance strategies remain unchanged.

8.2.2 $q = 1/2$

Again we begin by ignoring the option of silence, which we will discuss at the end of the subsection. Consider first acceptance decisions: Ll types offered 30, and Hh and Hl types offered 50.

(i) Consider first type Ll offered 30. The player knows that the opponent sent message h and will accept 70 regardless of type. Thus conditioning on acceptance offers no information. Taking into account $q = 1/2$:

$$\Pr(j \text{ is } L|(30,70),h_j) = \frac{1-\tau_L}{1-\tau_L+\tau_H}$$
$$\Pr(j \text{ is } H|(30,70),h_j) = \frac{\tau_H}{1-\tau_L+\tau_H}.$$

Ll accepts with positive probability if:

$$30\Pr(j \text{ is } H|(30,70),h_j) \geq 5\Pr(j \text{ is } L|(30,70),h_j)$$

or:

$$6\tau_H > 1-\tau_L \Longrightarrow \beta = 1,\ 6\tau_H < 1-\tau_L \Longrightarrow \beta = 0, \text{ and } 6\tau_H = 1-\tau_L \Longrightarrow \beta \in [0,1]. \tag{8}$$

(ii) Consider now type Hh, receiving recommendation $(50,50)$. Under the mediation mechanism, the player does not know the message sent by the opponent.

The relevant posterior probability is:

$$\Pr(j \text{ is } H \text{ and accepts } 50|(50,50),h_i) = \frac{\Pr(j \text{ is } H,(50,50),\ j \text{ accepts } 50|h_i)}{\Pr((50,50),|h_i)}$$

where:

$$\Pr(j \text{ is } H,(50,50),\ j \text{ accepts } 50|h_i) =$$
$$\Pr((50,50)|h_j,j \text{ is } H,h_i)\Pr(j \text{ is } H \text{ and accepts } 50|h_j,h_i)\Pr(h_j|j \text{ is } H)\Pr(H)+$$
$$\Pr((50,50)|l_j,j \text{ is } H,h_i)\Pr(j \text{ is } H \text{ and accepts } 50|l_j,h_i)\Pr(l_j|j \text{ is } H)\Pr(H)$$
$$= ((\tau_H/2)\alpha_h + (3/8)(1-\tau_H)\alpha_l)(1/2),$$

and:

$$\Pr((50,50)|h_i) = \Pr(j \textit{ is } H, (50,50)|h_i) + \Pr(j \textit{ is } L, (50,50)|h_i).$$

Substituting the relevant probabilities, and taking into account that L types always accept 50:

$$\Pr(j \textit{ is } H|(50,50),\ j \text{ accepts } 50, h_i) = \frac{4\tau_H\alpha_h + 3(1-\tau_H)\alpha_l}{4\tau_H + 3(1-\tau_H) + 4(1-\tau_L) + 3\tau_L}$$

and

$$\begin{aligned}\Pr(j \textit{ is } L|(50,50),\ j \text{ accepts } 50, h_i) &= 1 - \Pr(j \textit{ is } H|(50,50),\ j \text{ accepts } 50, h_i) \\ &= \frac{4(1-\tau_L) + 3\tau_L}{4\tau_H + 3(1-\tau_H) + 4(1-\tau_L) + 3\tau_L}.\end{aligned}$$

Hh will accept 50 with positive probability if:

$$15\Pr(j \textit{ is } H|(50,50),\ j \text{ accepts } 50, h_i) \geq 20\Pr(j \textit{ is } L|(50,50),\ j \text{ accepts } 50, h_i) \tag{9}$$

or:

$$\begin{aligned}15(4\tau_H\alpha_h + 3(1-\tau_H)\alpha_l) = 20(4(1-\tau_L) + 3\tau_L) &\Longrightarrow \alpha_h \in [0,1] \\ 15(4\tau_H\alpha_h + 3(1-\tau_H)\alpha_l) < 20(4(1-\tau_L) + 3\tau_L) &\Longrightarrow \alpha_h = 0, \\ 15(4\tau_H\alpha_h + 3(1-\tau_H)\alpha_l) > 20(4(1-\tau_L) + 3\tau_L) &\Longrightarrow \alpha_h = 1\end{aligned} \tag{10}$$

Condition (9) corresponds to (2) in the text, specialized to the experimental parameters.

(iii) Similarly, an H type who sent message l and is offered a $(50,50)$ split, will compute the posterior probability:

$$\Pr(j \textit{ is } H|(50,50),\ j \text{ accepts } 50, l_i) = \frac{3\tau_H\alpha_h + 8(1-\tau_H)\alpha_l}{3\tau_H + 8(1-\tau_H) + 3(1-\tau_L) + 8\tau_L}$$

and will accept 50 with positive probability if:

$$15\Pr(j \textit{ is } H|(50,50),\ j \text{ accepts } 50, l_i) \geq 20\Pr(j \textit{ is } L|(50,50),\ j \text{ accepts } 50, l_i)$$

or:

$$\begin{aligned} 15(3\tau_H\alpha_h + 8(1-\tau_H)\alpha_l) = 20(3(1-\tau_L) + 8\tau_L) &\implies \alpha_l \in [0,1] \\ 15(3\tau_H\alpha_h + 8(1-\tau_H)\alpha_l) < 20(3(1-\tau_L) + 8\tau_L) &\implies \alpha_l = 0 \\ 15(3\tau_H\alpha_h + 8(1-\tau_H)\alpha_l) > 20(3(1-\tau_L) + 8\tau_L) &\implies \alpha_l = 1 \end{aligned} \tag{11}$$

Conditions (8), (10), and (11) pin down the three probabilities β, α_h, and α_l as functions of τ_H and τ_L. Given these probabilities, the comparison of expected utilities at the message stage determines equilibrium τ_H and τ_L. If $\alpha_h = 1$, by Proposition 3, $\tau_H = 1$, $\tau_L = 1$. But if $\tau_H = 1$, then $\beta = 1$ by (8). The equilibrium in weakly undominated strategies then corresponds to the HMS equilibrium. Outside of such an equilibrium, $\alpha_h = 0$. Imposing $\alpha_h = 0$, the relevant expected utilities are:

$$\begin{aligned} EU_H(h) &= (1/2)[\tau_H 35 + (1-\tau_H)((5/8)35 + (3/8)(50\alpha_l + 35(1-\alpha_l)))] + (1/2)70 \\ EU_H(l) &= (1/2)[\tau_H 35 + (1-\tau_H)(50\alpha_l^2 + 35(1-\alpha_l^2))] + \\ &\quad (1/2)[\tau_L(\alpha_l 50 + (1-\alpha_l)70) + (1-\tau_L)((5/8)70 + (3/8)(50\alpha_l + 70(1-\alpha_l)))] \\ EU_L(l) &= (1/2)[\tau_H(5/8)30\beta + (1-\tau_H)50\alpha_l] + (1/2)[\tau_L 50 + (1-\tau_L)((5/8)(30\beta + 35(1-\beta)) + (3/8)50)] \\ EU_L(h) &= (1/2)[(1-\tau_H)((3/8)50\alpha_l] + \\ &\quad (1/2)[\tau_L((5/8)(70\beta + 35(1-\beta)) + (3/8)50) + (1-\tau_L)(50/2 + 35/2)] \end{aligned} \tag{12}$$

As before, four conditions, (11), (8), and the relevant expected utilities equations, determine β, α_l, τ_L and τ_H. One preliminary observation simplifies the identification of the equilibria:

Lemma A2. *If $q = 1/2$, there exist no equilibria for which $\alpha_l > 0$.*

Proof. The proof is in two steps. (1) Suppose first $\alpha_l \in (0,1)$. Then, from (11):

$$6(1-\tau_H)\alpha_l = 3 + 5\tau_L \implies \tau_H = 1 - \left(\frac{3+5\tau_L}{6\alpha_l}\right). \tag{13}$$

Substituting (13) in (12), we find that for any β:

$$EU_H(h) - EU_H(l) = (5/32)(3+5\tau_L)(3+5\alpha_l) > 0.$$

But then $\tau_H = 1$ and (13) is violated. Thus $\alpha_l \in (0,1)$ is impossible.[44]

(2) Suppose then $\alpha_l = 1$. From (11), it follows that:

$$\tau_H \leq (1/2) - (5/6)\tau_L. \tag{14}$$

Note that there cannot be an equilibrium with $\alpha_l = 1$ if L prefers sincerity and thus $\tau_L = 1$. From (12):

$$EU_L(l) - EU_L(h) = (5/16)[(47 - 5\beta) - \tau_H(50 + 30\beta) + \tau_L(18 - 30\beta)],$$

an expression that is minimal at τ_H maximum value. By (14), such maximal value must correspond to $\tau_H = (1/2) - (5/6)\tau_L$. Substituting, we then obtain:

$$EU_L(l) - EU_L(h) > 0 \Longleftarrow (5/48)[66 + 30\beta + \tau_L(179 - 165\beta)] > 0.$$

The condition is always satisfied. Hence $\tau_L = 1$; but then by (14) there cannot be an equilibrium with $\alpha_l = 1$, and the Lemma is proven. □

Proposition 3 and Lemma A2 establish $\alpha_l = 0$ and, unless $\tau_L = 1$ and $\tau_H = 1$, $\alpha_h = 0$. Studying (12) and (8), we can identify the full set of equilibria:[45]

$$(i)\ \alpha_h = 1, \beta = 1, \tau_L = 1, \tau_H = 1;$$

$$(ii)\ \alpha_h = 0, \beta = 1, \tau_L = 1, \tau_H = 1;$$

$$(iii)\ \alpha_l = 0, \alpha_h = 0, \tau_L = 0, \tau_H \leq 4/15;$$

$$(iv)\ \alpha_l = 0, \alpha_h = 0, \beta = 1, \tau_L \in (0,1), \tau_H = 4/15 + (6/15)\tau_L;$$

$$(v)\ \alpha_l = 0, \alpha_h = 0, \beta = 1, \tau_L = 1, \tau_H \in [2/3, 1);$$

$$(vi)\ \alpha_l = 0, \alpha_h = 0, \beta \in (0, 3/7),\ \tau_L = 3/(18 - 35\beta), \tau_H = (1/6)(1 - 3/(18 - 35\beta));$$

$$(vii)\ \alpha_h = 0, \beta = 0, \tau_L = 1/6, \tau_H \leq 5/36.$$

Equilibrium (i) is the HMS equilibrium.

[44] We are imposing $\alpha_h = 0$. But $\alpha_h = 1 \Longrightarrow (\tau_H = 1, \tau_L = 1)$. On the equilibrium path, α_l is irrelevant; off-equilibrium, by (11) an H player who lied would still reject 50.

[45] Equilibrium (iii) has message probabilities $\tau_L = 0$ and $\tau_H \leq 4/15$; if $\tau_H < 1/6$, the equilibrium is supported by the (sequentially rational) belief that were L to send message l and be offered 30, at the acceptance stage the offer would be rejected, if $\tau_H \in (1/6, 4/15]$, the equilibrium is supported by the rational belief that the offer would be accepted. At $\tau_H = 1/6$, either belief supports the equilibrium.

Silence As in the case of $q = 1/3$, with silence interpreted by the computer mediator according to the prior, the equilibria characterized above extend to the possibility of silent messages with a simple change of variable: τ_T becomes $\widehat{\tau}_T$ in all equations above and it is $\widehat{\tau}_T$ that is determined in equilibrium (that is, τ_T and σ_T are jointly determined).

Although the conclusion continues to hold, with $q = 1/2$, there is one complication: when messages are obfuscated by the mediator, a subject who sent a silent message will not know not only what message the opponent sent but also how the subject's own message was read by the computer. The reason this complication does not invalidate the previous analysis is that, in the absence of silence, equilibrium acceptance strategies depend only on type. More precisely, given the focus on equilibria in undominated strategies, the only acceptance strategies that could depend on the message sent are α_l and α_h.[46] But, barring full sincerity, $\alpha_l = \alpha_h = 0$ in all equilibria: H types reject 50 regardless of whether they sent message l or h. When silent messages are used, full sincerity is impossible, and for all $\widehat{\tau}_L$ and $\widehat{\tau}_H$ equilibria must exist where H types reject 50 regardless of how their message has been read by the computer. Hence, denoting by α_s the probability that an H type who sent a message s accepts 50, there must be equilibria with $\alpha_l = \alpha_h = \alpha_s = 0$: all H types reject 50. It follows that, substituting $\widehat{\tau}_T$ for τ_T, the equilibria described above remain equilibria when silent messages are possible.

8.3 Trembling-hand Perfection

As mentioned in the text, if $(2\theta - 1) < q < (2\theta - 1)/\theta$, the best equilibrium under optimal mediation is trembling-hand perfect (even while fragile in the sense of Proposition 3). Consider the following.

A perfect equilibrium cannot include weakly dominated strategies. Thus, if the equilibrium is perfect, all accept θ, L always accepts $1/2$, and H always rejects $(1 - \theta)$, all of which are in line with the HMS equilibrium. The L type ex-post participation constraint is slack in equilibrium; the three incentive constraints that bind in equilibrium and could be violated in the presence of trembles are the H type acceptance of $1/2$ following message h, the L type truthfulness constraint, and the H type truthfulness constraint including the possibility of double deviation (i.e. sending message l and then rejecting a recommendation of $1/2$). We write below the three conditions that must be satisfied for

[46] Recall that the recommendation $(70, 30)$ can only follow messages that have been read as (h, l). Hence there is no uncertainty on how one's own message (or for that matter, the opponent's) has been read. The possibility of silence affects the updating probability on the opponent's type and makes β a function of $\widehat{\tau}_L$, $\widehat{\tau}_H$. With this change in variable, the equilibrium conditions in (8) can be rewritten as before.

the prescribed strategies to be best responses, given trembles around equilibrium behavior. We then show that a vector of trembles exists such that all conditions are satisfied. Throughout we use the notation α_m^x (β_m^x) to denote the probability that an H (L) player who sent message m accepts the offer of x.

Consider first the acceptance strategy for a sincere H type who is offered 1/2 and in the HMS equilibrium accepts it. Call Hh player i, and j the opponent. Then:

$$EU_{Hh}(\text{accept } 1/2) \geq EU_{Hh}(\text{reject } 1/2) \iff$$

$$(1/2-\theta/2)\Pr(j \text{ accepts and is } H|h_i,(1/2,1/2)) \geq (\theta-1/2)\Pr(j \text{ accepts and is } L|h_i,(1/2,1/2))$$

or, borrowing from the proof of Proposition 3 in the text:

$$(1/2-\theta/2)q\left[q_H\tau_H\alpha_h^{1/2}+q_M(1-\tau_H)\alpha_l^{1/2}\right] \geq (\theta-1/2)(1-q)\left[q_H(1-\tau_L)\beta_h^{1/2}+q_M\tau_L\beta_l^{1/2}\right]$$

where, from (1) in the text:

$$q_M = \left(\frac{1-\theta}{2\theta-1}\right)\left(\frac{1+q-2\theta}{\theta-q}\right)$$
$$q_H = \left(\frac{1-q}{q}\right)\left(\frac{1+q-2\theta}{\theta-q}.\right)$$

Consider trembles such that: $\alpha_h^{1/2}=1-\varepsilon_{\alpha_h^{1/2}}$, $\alpha_l^{1/2}=\varepsilon_{\alpha_l^{1/2}}$, $\beta_h^{1/2}=1-\varepsilon_{\beta_h^{1/2}}$, $\beta_l^{1/2}=1-\varepsilon_{\beta_l^{1/2}}$, $\tau_H=1-\varepsilon_{\tau_H}$, $\tau_L=1-\varepsilon_{\tau_L}$. We want to know whether there exist a vector $\varepsilon_1=\{\varepsilon_{\alpha_h^{1/2}}, \varepsilon_{\alpha_l^{1/2}}, \varepsilon_{\beta_h^{1/2}}, \varepsilon_{\beta_l^{1/2}}, \varepsilon_{\tau_H}, \varepsilon_{\tau_L}\}$ such that:

$$\lim_{\varepsilon_1 \longrightarrow 0}\Big[(1/2-\theta/2)q\left(q_H(1-\varepsilon_{\tau_H})(1-\varepsilon_{\alpha_h^{1/2}})+q_M\varepsilon_{\tau_H}\varepsilon_{\alpha_l^{1/2}}\right) -$$
$$(\theta-1/2)(1-q)\left(q_H\varepsilon_{\tau_L}(1-\varepsilon_{\beta_h^{1/2}})+q_M(1-\varepsilon_{\tau_L})(1-\varepsilon_{\beta_l^{1/2}})\right)\Big] \geq 0$$

Set $\varepsilon_{\alpha_h^{1/2}}=a_h^{1/2}/n$, $\varepsilon_{\tau_H}=t_H/n$, $\varepsilon_{\alpha_l^{1/2}}=a_l^{1/2}/n$, $\varepsilon_{\beta_h^{1/2}}=b_h^{1/2}/n$, $\varepsilon_{\tau_L}=t_L/n$, $\varepsilon_{\beta_l^{1/2}}=b_l^{1/2}/n$, where

$\{a_h^{1/2}, t_H, a_l^{1/2}, b_h^{1/2}, t_L, b_l^{1/2}\}$ is a vector of positive constants. The condition becomes:

$$\lim_{n\longrightarrow\infty}\Big[(1/2-\theta/2)q\left(q_H(1-t_H/n)(1-a_h^{1/2}/n)+q_M(t_H/n)(a_l^{1/2}/n)\right)- \\ (\theta-1/2)(1-q)\left(q_H(t_L/n)(1-b_h^{1/2}/n)+q_M(1-t_L/n)(1-b_l^{1/2}/n)\right)\Big]\geq 0 \quad (15)$$

We also need to verify that there exists trembles such that both types prefer to be truthful. For a player of type L we require $EU_L(l)\geq EU_L(h)$ where:

$$EU_L(l)=q(\tau_H[(1-q_M)(1-\theta)\alpha_h^{\theta}\beta_l^{1-\theta}+q_M(1/2)\alpha_h^{1/2}\beta_l^{1/2}]+(1-\tau_H)[1/2)\alpha_l^{1/2}\beta_l^{1/2}])+ \\ (1-q)(\tau_L[(1/2)(\beta_l^{1/2})^2+(\theta/2)(1-(\beta_l^{1/2})^2)]+ \\ (1-\tau_L)[(1-q_M)((1-\theta)\beta_l^{1-\theta}\beta_h^{\theta}+(\theta/2)(1-\beta_l^{1-\theta}\beta_h^{\theta}))+q_M((1/2)\beta_l^{1/2}\beta_h^{1/2}+\theta/2(1-\beta_l^{1/2}\beta_h^{1/2}))])$$

and:

$$EU_L(h)=q(\tau_H[q_H(1/2)\alpha_h^{1/2}\beta_h^{1/2}]+(1-\tau_H)[(1-q_M)(\theta\alpha_l^{1-\theta}\beta_h^{\theta})+q_M(1/2)\alpha_l^{1/2}\beta_h^{1/2}]+ \\ (1-q)(\tau_L[(1-q_M)(\theta\beta_l^{1-\theta}\beta_h^{\theta}+(\theta/2)(1-\beta_l^{1-\theta}\beta_h^{\theta}))+q_M((1/2)\beta_l^{1/2}\beta_h^{1/2}+(\theta/2)(1-\beta_l^{1/2}\beta_h^{1/2}))]+ \\ (1-\tau_L)[q_H((1/2)(\beta_h^{1/2})^2+(\theta/2)(1-(\beta_h^{1/2})^2))+(1-q_H)(\theta/2)].$$

For a player of type H, we require $EU_H(h)\geq EU_H(l)$ where:

$$EU_H(h)=q(\tau_H[(1-q_H)(\theta/2)+q_H((1/2)(\alpha_h^{1/2})^2+(\theta/2)(1-(\alpha_h^{1/2})^2)]+(1-\tau_H)[q_M((1/2)\alpha_h^{1/2}\alpha_l^{1/2}+ \\ (\theta/2)(1-\alpha_h^{1/2}\alpha_l^{1/2}))+(1-q_M)(\theta\alpha_h^{\theta}\alpha_l^{1-\theta}+(\theta/2)(1-\alpha_h^{\theta}\alpha_l^{1-\theta}))])+ \\ (1-q)(\tau_L[(1-q_M)\theta+q_M((1/2)\alpha_h^{1/2}\beta_l^{1/2}+\theta(1-\alpha_h^{1/2}\beta_l^{1/2}))]+ \\ (1-\tau_L)[(1-q_H)\theta+q_H((1/2)\alpha_h^{1/2}\beta_h^{1/2}+\theta(1-\alpha_h^{1/2}\beta_h^{1/2}))])$$

and:

$$EU_H(l) = q(\tau_H[(1-q_M)((1-\theta)\alpha_h^{\theta}\alpha_l^{1-\theta} + (\theta/2)(1-\alpha_h^{\theta}\alpha_l^{1-\theta}))+$$
$$q_M((1/2)\alpha_h^{1/2}\alpha_l^{1/2} + (\theta/2)(1-\alpha_h^{1/2}\alpha_l^{1/2}))] + (1-\tau_H)[(1/2)(\alpha_l^{1/2})^2 + (\theta/2)(1-(\alpha_l^{1/2})^2)])+$$
$$(1-q)(\tau_L[(1/2)\alpha_l^{1/2}\beta_l^{1/2} + \theta(1-\alpha_l^{1/2}\beta_l^{1/2})] + (1-\tau_L)[(1-q_M)((1-\theta)\alpha_l^{1-\theta}\beta_h^{\theta} + \theta(1-\alpha_l^{1-\theta}\beta_h^{\theta}))+$$
$$q_M((1/2)\alpha_l^{1/2}\beta_h^{1/2} + \theta(1-\alpha_l^{1/2}\beta_h^{1/2}))]).$$

As above, set $\alpha_h^{\theta} = 1 - a_h^{\theta}/n$, $\alpha_h^{1/2} = 1 - a_h^{1/2}/n$, $\alpha_l^{1/2} = a_l^{1/2}/n$, $\alpha_l^{1-\theta} = a_l^{1-\theta}/n$, $\beta_h^{\theta} = 1 - b_h^{\theta}/n$, $\beta_h^{1/2} = 1 - b_h^{1/2}/n$, $\beta_l^{1/2} = 1 - b_l^{1/2}/n$, $\beta_l^{1-\theta} = 1 - b_l^{1-\theta}/n$, and recall $\tau_H = 1 - t_H/n$, $\tau_L = 1 - t_L/n$. We want to verify that there exist a vector of positive constants $\{t_H, t_L, a_h^{1/2}, a_h^{\theta}, a_l^{1/2}, a_l^{1-\theta}, b_l^{1/2}, b_l^{1-\theta}, b_h^{1/2}, b_h^{\theta}\}$ such that (15) is satisfied, as well as:

$$\lim_{n \longrightarrow \infty} [EU_L(l) - EU_L(h)] \geq 0$$

and:

$$\lim_{n \longrightarrow \infty} [EU_H(h) - EU_H(l)] \geq 0.$$

It is not difficult to find a vector that satisfies the conditions at the experimental parameter values if $q = 1/2$ and $\theta = 0.7$, as in our experimental parameterization. For example, all three conditions are satisfied at $\{t_H = 1, t_L = 1, a_h^{50} = 1, a_h^{70} = 1, a_l^{50} = 1, a_l^{30} = 1, b_l^{50} = 4, b_l^{30} = 8, b_h^{50} = 8, b_h^{70} = 4\}$. The equilibrium is perfect as long as beliefs assign higher probability to L types' trembles that result in rejections, and low probability to deviations from truthfulness and to trembles in H's acceptance strategies. The result is not surprising: as Proposition 3 leads us to expect, the condition most difficult to satisfy is the acceptance of offer 50 by H types in the presence of noise in behavior. Even in the presence of noise, however, acceptance is a best response if the probability of deviations from truthfulness and from other H's accepting is low, relative to the probability of rejections by L types. The positive conclusion reflects the latitude in the choice of trembles.

The result is not limited to the experimental parameter values. We have verified numerically that the vector of trembles identified above support trembling-hand perfection for arbitrary q and θ in the interval $(2\theta - 1) < q < (2\theta - 1)/\theta$, with $\theta/2 > 1 - \theta$.