Mediating Conflict in the Lab[*]

Alessandra Casella[†] Evan Friedman[‡] Manuel Perez Archila[§]

March 14, 2022

Abstract

Mediation of disputes is increasingly common, and computer-run algorithms are taking a central role. We test the efficacy of a theoretically optimal mediation algorithm in an experiment where two subjects negotiate how to share a resource. The subjects send cheap talk messages to one another (under direct communication) or to the computer mediator (under mediation), before expressing demands or receiving the mediator's recommendation. While messages to the computer mediator are more sincere, peaceful resolution is not more frequent. We argue, theoretically and experimentally, that exactly when it is most promising, the mechanism is fragile to any deviation from full sincerity.

1 Introduction

From family disputes to labor relations to corporate law, mediation attracts sustained, indeed increasing, attention, from psychologists, lawyers and judges, lay people, and professional agents specializing in its craft. Alternative Dispute Resolution procedures have become favored solutions to the time delay and costs of judicial resolutions and have florished particularly online, a side-effect of the growth of e-commerce and the development of smarter algorithms.[1]

Yet, the potential of mediation to facilitate the resolution of conflict remains intriguing. In its purest form, and as studied in this paper, the mediator is an impartial third party who has neither independent resources, nor superior information, nor enforcement power. On what basis can the

[*] 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

[§] Princeton University, mp1278@princeton.edu

[1] The number of legal cases brought to trial in US courts experienced a startling 60% decline from the mid 80's to 2002 (Galanter, 2004). For useful entries into Alternative Dispute Resolution procedures, see Lodder and Zeleznikow (2010) and Barnett and Treleaven (2018). Wikipedia's page on online dispute resolution (https://en.wikipedia.org/wiki/Online_dispute_resolution, accessed February 14, 2022) offers a panoramic view.

mediator's presence be helpful? Part of the answer may be psychological: the presence of the mediator may prevent escalation when emotions run high. But mechanism design teaches us that mediation can help even when parties' interactions are coldly rational. The essence is the confidentiality of the communication between the parties and the mediator. It is possible for the mediator to induce the parties to reveal their private information, and yet issue recommendations that leave them uncertain about their opponent and thus willing to accept the mediator's recommendation. The final result is a higher frequency of peaceful resolutions than what the two sides can obtain by communicating directly. As Roger Myerson phrases it, the key is the *obfuscation* the mediator can employ: by leaving each side uncertain about the information disclosed by the opponent, the mediator can reach agreements that are impossible under direct communication (Myerson, 1991, ch.6). The importance of confidentiality is readily recognized by practitioners: having separate communication channels with each party is considered an essential ingredient of successful mediation, and one that deserves and requires protection (American Bar Association, 2005).

In this paper, we bring to the lab the theoretical model of mediation in Hörner, Morelli and Squintani (2015). We test the optimal mechanism identified by Hörner et al. by embodying it in an algorithm that issues recommendations to the experiment's participants, as parties in a dispute. We thus offer a test of a specific form of (theoretically optimal) algorithmic mediation. Our intended contribution is on two fronts: as an experimental test of a seminal theoretical result, and as a step towards the rigorous analysis of mediation algorithms.

In the model and in the experiment, two players negotiate how to share a resource. In case of conflict, the players' privately known strengths determine their payoffs. The players send cheap talk messages about their strengths 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 recommendation. Under mediation, a commonly known algorithm responds to the players' messages, either issuing a recommendation or refusing to mediate. Peace prevails if a recommendation is made and both players accept it.

Existing mediation algorithms vary greatly in their scope of application and in their design. Beyond mechanisms devoted to e-commerce disputes, among the best known, and closest to us, are algorithms that help allocate resources between two parties with conflicting claims. Particularly when applied to disputes stemming from a divorce, they aim at fair and envy-free outcomes by inducing parties to reveal their priorities.[2] In other applications, the algorithms employ systems of blind bidding, where the parties repeatedly and privately adjust their reservation bids until an area of agreement opens up.[3] As in the mechanism we study, the algorithms stress the confidentiality of the parties' messages. In contrast to our mechanism, however, the simpler systems devote less explicit attention to the distribution of resources that would arise if mediation fails. The core of the problem we study is the revelation of the parties' chances of prevailing in case of conflict, as opposed to their private valuations for heterogeneous goods.

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, in our experiment mediation does not increase the frequency of peace: in the lab, mediation does not fulfill its promise.

Having established this result, we devote the remainder of the paper to understanding its causes. The experimental data lead us to our most novel theoretical finding: the fragility of the obfuscation mechanism. Whereas the optimal equilibrium involves full sincerity of messages, under obfuscation, equilibria with even the smallest deviation from full sincerity imply a discontinuous jump downward in the probability of peace of a sufficient magnitude to undo the advantages of mediation. There are parameter values for which the optimal mechanism is robust to small deviations from full truthfulness, but this can occur only if optimal mediation does not involve obfuscation and the mediator's recommendations reveal the parties' messages. The problem is that the theoretical superiority of mediation relies on obfuscation. Whenever optimal mediation involves obfuscation, the optimal equilibrium improves over direct communication, but is fragile. Whenever optimal mediation does not involve obfuscation, the optimal equilibrium is robust, but does not bring more frequent peace than can be achieved by direct communication.

The result is interesting for two reasons. The first is specific to conflict mediation. The vulnerability of the equilibrium with obfuscation does not appear to have been noticed in the literature, and the finding can matter for applications. For example, Meirowitz et al. (2019), again working with the Hörner et al. model, singles out the mediation mechanism with obfuscation as the one dispute resolution institution for international conflicts that could discourage increased militarization. Our analysis invites some caution. The second reason is broader. Obfuscation in mediating conflict is one

[2] See in particular the Adjusted Winner procedure (Brams and Taylor, 1996), applied commercially to dispute resolutions by https://www.fairproposals.com.

[3] See for example: //www.smartsettle.com/, in particular the simpler SmartsettleONE system.

example of the use of randomization in constructing optimal mechanisms in cheap talk communication. Applications range from third-party garbling in Sender-Receiver games of cheap talk (Myerson (1982), Blume et al. (2007)) to whistle-blowing (Chassang and Padro' i Miguel, 2019) to survey design (Warner, 1965). Rigorous experiments are few, but the available results are consistent. Optimal mechanisms with randomizations fall short: while they increase sincerity, they do not induce full truthfulness, and lead to outcomes that are broadly comparable to simpler direct elicitation (John et al., 2018, Blume 2019a, Blume 2019b). Within the problem we study, we document a very similar result and provide a rigorous justification by identifying the discontinuity in equilibrium payoffs that, under obfuscation, must accompany small deviations from truthful messages.

If the fragility of the obfuscation equilibrium is our most novel finding, the experiment shows that the optimal mechanism has also other, better known vulnerabilities. Multiplicity of equilibria is a well-known problem in mechanism design, and not surprisingly the lab makes it salient.[4] Noise in subjects' behavior, although consistently small enough to approximate best responses in individual behavior, nevertheless increases the frequency of conflict.[5]

Our study contributes to the literature on 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 comparison 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. The theoretical literature is both large and sophisticated. With the exception of Blume et al. (2019b), however, rigorous experimental tests of mediation are lacking.[6]

Beyond the specific focus on mediation, our work tests the ability of experimental participants to

[4] See, for example, Palfrey (1990) for a theoretical dicussion, and Cason et al. (2006), where multiplicity hampers the implementation of a desirable social choice, for impact on experimental results.

[5] 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.

[6] Experimental work on mediation in Political Science is less tied to theory and closer to historical events. For example, Wilkenfeld et al. (2003) simulate historical world crises and observe the impact of a mediator, trained to follow different protocols.

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). These works have found that the information design problem is particularly challenging for inexperienced subjects. Embodying the optimal mediation mechanism in a public algorithm simplifies the problem by guaranteeing that the mediation program is commonly known and committed to.

Our study is also close to the tradition of experiments in applied 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.[7] From an 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 with some positive probability. Is this a reasonable assumption? When mediation takes the form of an algorithm, commitment is built into the system. But even in the case of professional human mediators, with long term reputations to preserve, the ability to commit to refuse mediation can be a realistic assumption. 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. This said, in part to understand better the empirical importance of commitment, in part to explore potential biases of a human mediator, the experiment also includes an exploratory treatment where the mediator is played by one of the participants. Interestingly, conflict seems to be induced more by biases in the parties' behavior than by the mediator's lack of commitment.

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 algorithm is not more successful than direct communication in averting conflict in the lab. Section 6 presents briefly the results of the "human mediator" treatment. Finally, Section 7 concludes.

[7] 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.

2 The Model

The mediation game we took to the lab follows closely the model in Hörner, Morelli and Squintani (2015) (HMS). 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 $q < (2\theta - 1)/\theta$.

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. 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.[8] We assume $\theta/2 > 1 - \theta$, to ensure that the H type prefers to fight rather than to accept the smaller share when facing another H type.

[8] 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.

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, we constrain $r(m_1, m_2)$ 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 makes 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 refuses 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 refuse to mediate with positive probability induces players to 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.*

Mediation nests direct communication as a special case, so it is obvious that optimal mediation must result in a weakly higher probability of peace than unmediated communication. But HMS' result is stronger: for parameters in the specified range, mediation can achieve a *strictly* higher probability.

By the revelation principle (Myerson, 1982), there is an optimal mediation program that is also a direct revelation mechanism. Under the optimal such program, there exists an equilibrium where all messages to the mediator reveal the players' types sincerely, and the 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, following h messages, the mediator refuses to mediate with positive probability, thus keeping L sincere–the mediator is able to *commit* to refusing mediation. 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 to what can be achieved under unmediated communication, where truthful messages necessarily reveal the opponents' type.[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 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.

In characterizing the optimal equilibrium under direct communication, HMS allow for a publicly observed correlation device. The equilibrium involves a positive probability of war following specific pairs of messages, mimicking the commitment demanded from the mediator. In the lab, in the absence of repetition or external aids to commitment, achieving such an equilibrium is very difficult in principle and, we believe, impossible in practice.[10] For this reason, Proposition HMS is even more likely to hold in the lab, and we derive additional predictions without positing any correlation device. Specifically, we concentrate on Perfect Bayesian Equilibria (PBE) in undominated strategies that are symmetric for players of a given type. We refer to them in short as "equilibria", and we apply the same concept for all communication protocols we consider throughout the paper.

For the range of θ that we study, there is no equilibrium with full sincerity in the unmediated communication game. Intuitively, only the threat of war can induce sincerity, but walking out ($d = w$) is weakly dominated by making a large demand ($d = \theta$).

Proposition 1. *If $\theta/2 > 1 - \theta$, players' types cannot be fully revealed in any equilibrium of the unmediated communication game.*

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 $1/2$ and θ. Now consider a match between an H and an L. In such a match, the H player can always guarantee itself

[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).

[10] In the lab there is no explicit public correlation device and the optimal equilibrium cannot be achieved through the randomization of individual messages (because mixed strategy profiles cannot typically result in correlated randomness (Forges, 1986), as also noted by HMS).

[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.

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 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. □

The contrast between the optimal equilibrium under direct communication in HMS and Proposition 1 hinges on the implicit potential for commitment to walking out. We can also ask about the role of commitment under mediation. In the experiment, optimal mediation (with commitment) is implemented by a computer algorithm, but we also investigated a more exploratory treatment where the role of mediator was played by experimental subjects. With anonymity and random rematching across rounds, mediators (as well as players) cannot build reputation, and so there is no incentive to walk out of mediation. The absence of commitment power thus 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.*[12]

In the online Appendix (Sections 9.2 and 9.3), we characterize equilibria of the human mediator game and of the direct communication 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.

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

In taking the HMS model to the lab, we made two modifications, neither of which affects the model's theoretical properties. First, as described, we allow players to send a silent message. Silence

[12] Recall that $q < (2\theta-1)/\theta$ is a maintained assumption both in HMS and here.

[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, but again restricting attention to truthful revelation mechanisms.

does not affect the predictions about optimal mediation, but gives inexperienced subjects the intuitive option of hiding their type while they learn the game, without complicating the data with random exploratory messages.[14] Second, we constrain both demands and the mediator's recommendations to lie in a restricted set containing only those that appear in the HMS optimal equilibria, which simplifies the subjects' problem without affecting equilibria.

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 algorithm's recommendations are the following:[15]

$\underline{q = 1/2}$. *$r(l,l) = (0.5, 0.5)$; $r(h,l) = \{(0.7, 0.3)$ 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 $7/9 \approx 0.778$.*

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 optimal program 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 involves obfuscation. An H type recommended $(0.5, 0.5)$ prefers peaceful resolution when faced with another H type, but would rather go to war with an L type. In the optimal equilibrium with sincere players, obfuscation makes it so that the H type offered $(0.5, 0.5)$–uncertain over the opponent's type–is just indifferent and accepts as part of the equilibrium.[16] 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: after an H type sends message h, either the program 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 com-

[14] We thank Johannes Hörner for suggesting the modification.

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

[16] Similarly, in the optimal equilibrium with sincere players, an L type offered (0.5, 0.5) does not know her opponent's type. However, this uncertainty is less relevant for the L type, for whom accepting 0.5 is weakly dominant and a strict best response in the optimal equilibrium.

plicates the subjects' problem. Collecting data under both $q = 1/2$ and $q = 1/3$ allows us to study how subjects react to the optimal mediation programs in the two cases.

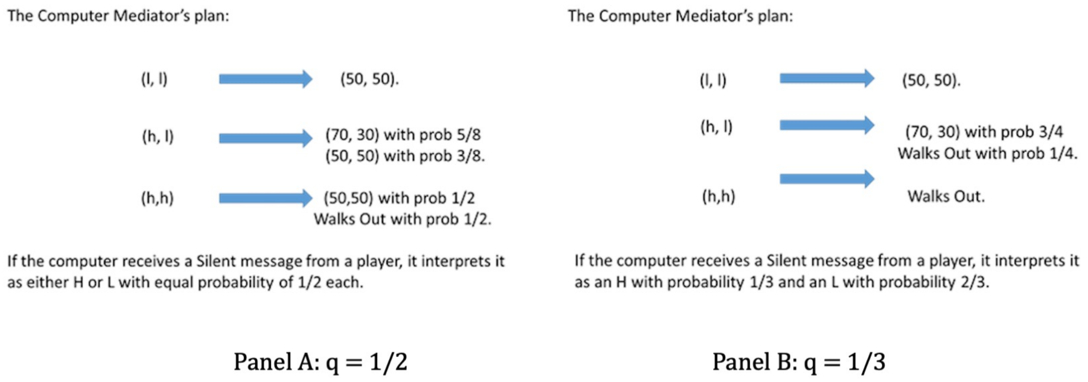
We ran the experiment at Columbia's Experimental Lab for the Social Sciences (CELSS) with subjects recruited through the lab's ORSEE recruitment system (Greiner, 2015). Most subjects were undergraduate students at Columbia University and Barnard College. The experiment lasted about 90 minutes and earnings 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 supplemental material (Section 10.4), where the instructions for one of the treatments are also reproduced.[17]

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. At the end of each round, 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.

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



The Computer Mediator's plan:

(l, l) → (50, 50).

(h, l) → (70, 30) with prob 5/8
(50, 50) with prob 3/8.

(h,h) → (50,50) with prob 1/2
Walks Out with prob 1/2.

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

Panel A: q = 1/2

The Computer Mediator's plan:

(l, l) → (50, 50).

(h, l) → (70, 30) with prob 3/4
Walks Out with prob 1/4.

(h,h) → Walks Out.

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

Panel B: q = 1/3

Figure 1: The Computer Mediator program as shown to subjects

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). As projected on the screen and emphasized during instructions, the computer interpreted silence according to the prior: as h with probability q and as l with probability $1 - q$.

Unless the computer chose w, the recommendation was conveyed to each subject who then chose, separately, either to accept it or reject it. The computer's recommendation 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, the two players 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.[18] 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.[19] Proposition 2 continues to apply to our HM game.

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.[20]

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:[21]

Experimental Design

[18] 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 $(1-q)$). In the supplemental material (Section 10.2), we show that, when $q = 1/3$, there are no non-pooling equilibria in which H is truthful, HM offers $(50, 50)$ w.p. 1 following (l, l) and $(70, 30)$ with positive probability following (h, l).

[19] Note that the numerical values for the mediator's payoffs were kept constant across the two parametrizations. In a different setting from ours, Ivanov (2010) develops a framework for eliciting an optimal mediation program from a strategic mediator.

[20] The 30 rounds allowed participants to develop some experience as mediator and to generate 20 rounds of data as disputing parties, as in the other treatments.

[21] When CM was ordered first among the treatments of interest, under Order 2, we added two more practice rounds with CM, to aid subjects' understanding of the less intuitive game. To limit the number of sessions, we also kept constant the order of the HM treatment, at an intermediate position that results in a level of experience comparable on average to CM and UC.

Sessions	q	Order	Subjects	Groups / Treatment	Rounds	Groups × Rounds
s1-s3	1/2	1: NC,UC,HM,CM	12×3	$6, 6, 4, 6$	$10, 20, 30, 20$	$60, 120, 120, 120$
s4-s6	1/3	1: NC,UC,HM,CM	12×3	$6, 6, 4, 6$	$10, 20, 30, 20$	$60, 120, 120, 120$
s7-s9	1/2	2: NC,CM,HM,UC	12×3	$6, 6, 4, 6$	$10, 20, 30, 20$	$60, 120, 120, 120$
s10-s12	1/3	2: NC,CM,HM,UC	12×3	$6, 6, 4, 6$	$10, 20, 30, 20$	$60, 120, 120, 120$

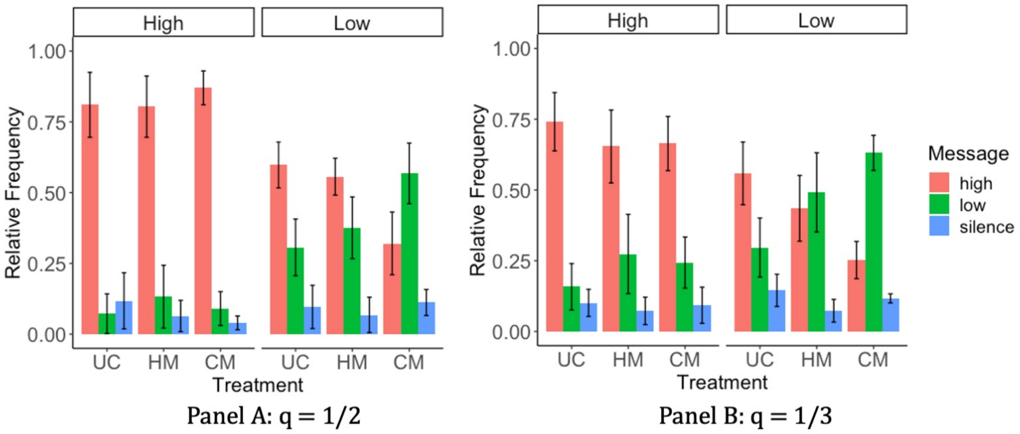
Because we always ordered NC first and, as mentioned, treated it as practice, we do not compare its results to the other treatments and do not discuss it in the text. For completeness, we describe the NC data as well as the equilibria of the NC game in the supplemental material (Section 10.3.4). Using data from Order 1 and Order 2, the design also allows us to compare UC and CM between subjects, when each of the two treatments is run immediately after NC. In the supplemental material (Section 10.3.2), we replicate the results discussed below over this restricted data set.

4 Experimental Results

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 session level.

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 $q = 1/3$. For each value of q, there is no detectable treatment effect for H types. 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



Panel A: q = 1/2

Panel B: q = 1/3

Figure 2: Messages by type and treatment

from 31 percent in UC to 57 percent in CM if $q = 1/2$, and from 30 to 64 percent if $q = 1/3$. F tests confirm that this difference is statistically significant at the 1 percent level for both parameters. 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.

As shown in the first column of Table 1, a linear probability model confirms what the figures show.[22] L types are less sincere than H types, and for L types treatment effects are present and 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.[23]

Finally, as shown in the second column of Table 1, the use of silence is not only scarce but declines with experience. Although an intuitive choice, silence is a redundant option, and one that should not be used in the optimal equilibria. Its decline with experience is a useful check on subjects' attention and understanding of the game.

[22] We report all regression results in the paper as estimated from a linear probability model; in all cases we have verified that qualitative results are unchanged under probit. We also verified that the results are unchanged when clustering errors at the individual level (for messaging) and at the pair of subjects level (for peace).

[23] Note that sincerity does not map directly into the information conveyed. How much a message moves the posterior probability of a given type depends on the use of the message by both types. In the supplemental material (Section 10.3.1), we report Kullback Leibler (KL) measures applied to our data. It remains true that the treatment conveying most information is CM, for both messages and in both parameterizations.

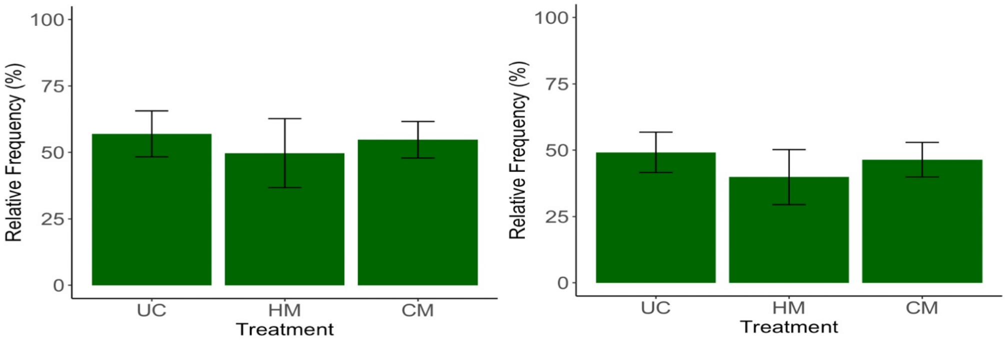
	Dependent variable:	
	Sincerity	Silence
	(1)	(2)
HM Treatment	−0.040 (0.040)	−0.042 (0.027)
CM Treatment	0.005 (0.048)	−0.049 (0.024)
Order 2	0.009 (0.047)	−0.068 (0.023)
$q = 1/2$	0.142 (0.046)	−0.015 (0.025)
L-type	−0.356 (0.136)	−0.029 (0.046)
Round	0.002 (0.001)	−0.001 (0.000)
HM treatment × L-type	0.181 (0.070)	−0.013 (0.028)
CM treatment × L-type	0.298 (0.074)	0.038 (0.034)
Order 2 × L-type	−0.025 (0.086)	0.017 (0.012)
$q = 1/2$ × L-type	−0.198 (0.085)	−0.005 (0.018)
Round × L-type	−0.0001 (0.001)	0.001 (0.001)
Constant	0.611 (0.071)	0.212 (0.040)
Observations	8,640	8,640

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

Table 1: Sincerity and Silence.

4.2 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 session level). Whether $q = 1/2$ or $q = 1/3$, there is no significant difference across treatments. In both cases, treatment HM results



Panel A: q = 1/2

Panel B: q = 1/3

Figure 3: Frequency of peace.

in least frequent peace and UC in most frequent peace, but the effects are small.[24] 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 supplemental material (Section 10.3.3).

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, 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 it increases significantly but very little over time. Across treatments, UC and CM are comparable, while peace is significantly lower under HM.[25]

5 Mediation Increases Sincerity but not Peace. Why?

Theory gives tighter 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

[24] 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 $q = 1/3$.

[25] In the supplemental material (Section 10.3.2), Figure 16 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 supplemental material (Section 10.3.3) 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.

	Dependent variable:	
	Peace	
	(1)	(2)
HM Treatment	-0.082	-0.085
	(0.039)	(0.039)
CM Treatment	-0.025	-0.018
	(0.032)	(0.027)
Order 2	0.004	0.011
	(0.046)	(0.047)
$q = 1/2$	0.087	0.186
	(0.046)	(0.044)
H-L pair		0.293
		(0.028)
L-L pair		0.606
		(0.029)
Round	0.001	0.001
	(0.0005)	(0.0005)
Constant	0.436	0.042
	(0.049)	(0.052)
Observations	4,320	4,320

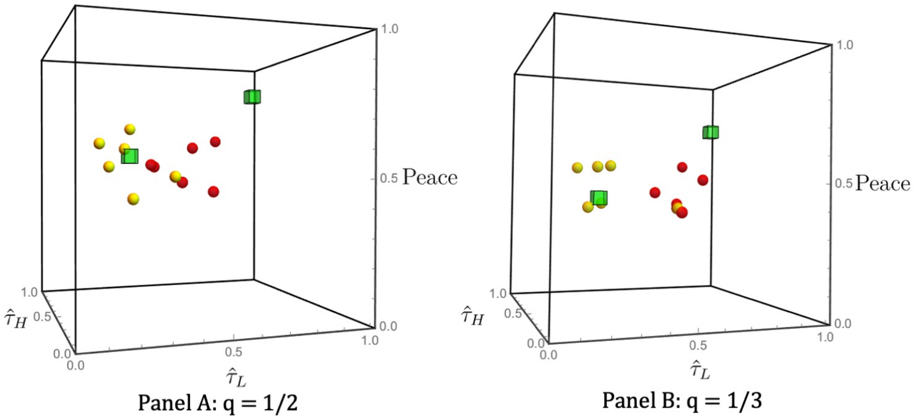
The default treatment is UC, Order 1, $q = 1/3$, and when looking at different pair types, the default pair is H-H. Standard errors are clustered at the session level.

Table 2: Peace.

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 $\widehat{\tau}_H = \tau_H + q\sigma_H$ as the frequency of all messages sent by H subjects that are read as h by the computer. Because it is informative to compare visually the results from CM and UC, for the purposes of this figure only, we also code silent messages in UC according to the prior (i.e. as h with frequency q). We have verified that, because silence is rarely used, the plot hardly changes under any other possible imputation.

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[26] (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).



Panel A: q = 1/2

Panel B: q = 1/3

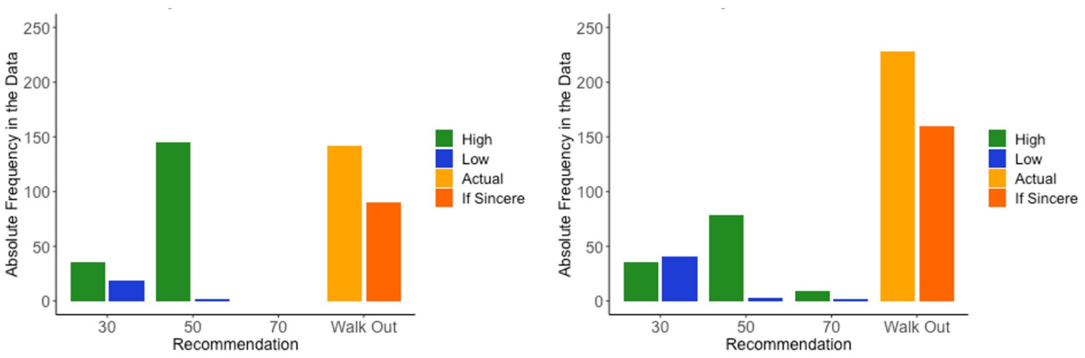
Figure 4: Sincerity and peace in UC and CM sessions.

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, 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. To better visualize the data, the reader may view animations of Figure 4 and other 3D figures that appear later at our website (http://www.evankfriedman.com/mediation.html).

Why wasn't the promise of optimal mediation realized in the data? Figure 5 gives some indications 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.[27] In the optimal equilibrium, all messages are sincere, all offers are accepted, and conflict

[26] For UC, the green cube corresponds to the equilibrium closest to the data in the $\{\tau_L, \tau_H, peace\}$ plane among equilibria characterized in the online Appendix, Section 9.3.

[27] The figures report individual rejections of offers. Because a single rejection is sufficient to trigger conflict, there can



Panel A: q = 1/2

Panel B: q = 1/3

Figure 5: Causes of war.

Green and blue columns correspond to rejections of recommendations, by H and L types respectively. Light and dark orange columns correspond to CM's refusals to mediate, in the data and if players had been sincere respectively.

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.

With both parameterizations, dominated actions (L rejecting 50, either type rejecting 70) are rare. When $q = 1/2$, excess conflict 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's frequency of lies (see Figure 2).

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. It is the case that the theoretical success of mediation is predicated on an equilibrium with full sincerity?

be some double counting: two individual rejections can amount to a single offer being turned down.

5.1 Multiple Equilibria and the Fragility of Peace under Obfuscation

Keeping fixed the mediator's program, we study the equilibria of the CM treatment.[28] We concentrate on equilibria in undominated strategies where, regardless of message: (i) all players accept 70; (ii) L players accept 50; (iii) H players 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 player accepting 50, and by β the probability of an Ll player accepting 30 (the offer of 30 can only follow an l message). As before, we denote by $\widehat{\tau}_T$ the probability that the message of type $T \in \{H, L\}$ is read as T by the algorithm (including the option of silence).

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

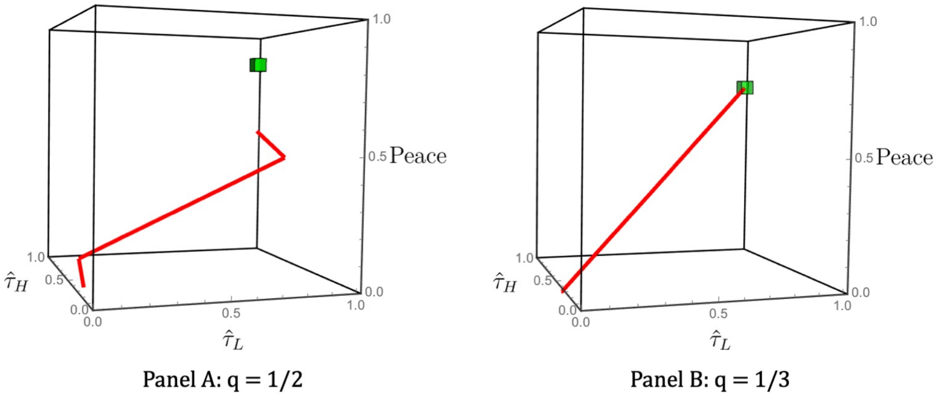
Table 3. Equilibria under CM

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.[29]

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 values of q, there are equilibria supporting a large range of peace probabilities, and any frequency of truthfulness, from

[28] Because our objective here is to understand 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 $\theta/2 > (1 - \theta)$.

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



Panel A: q = 1/2

Panel B: q = 1/3

Figure 6: Equilibria under CM

0 to 1, for L's and almost as large a range for H's. Keeping fixed the mediator's program, equilibrium behavior under CM is compatible with a large range of messages and outcomes.

Beyond depicting such wide variation, the most striking feature of the figure is the discontinuity in the locus of equilibria under $q = 1/2$. 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 best equilibrium under obfuscation is fragile.

In Proposition 3 below, we show that the discontinuity does not depend on the specific parameters used in the experiment; it 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 broader theoretical interest, beyond the specific results of our experiment. We phrase the proposition for generic parameter values in the appropriate range, and in the language of HMS, ignoring the option of silent messages.[30] (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$\}$;

[30] 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.

$r(h,h) = \{(1/2,1/2)$ with probability q_H and w otherwise$\}$ where:

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

Again using α_h (α_l) for the probability that Hh (Hl) accepts 1/2, we have the following proposition:

Proposition 3. *Suppose* $(2\theta-1) < q < (2\theta-1)/\theta$. *Then, in equilibrium: (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)* $\{\tau_H = 1, \tau_L = 1\} \not\Rightarrow \alpha_h = 1$.

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:

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

Using Bayes' rule gives:

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

and similar expressions for $Pr(Hl_j \mid (1/2,1/2),h_i)$ and $Pr(L_j|(1/2,1/2),h_i)$.[31] Using these and (1), $\alpha_h > 0$ 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 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 most likely to be satisfied at these values, at which it simplifies to the equality $(1-q) = (1-q)$. 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

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

induces the Hh player to *always* reject $1/2$. In fact, an equilibrium where Hh rejects $1/2$ exists even with full sincerity. 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.

Surprisingly, the equilibrium with $\alpha_h = 1$ is trembling-hand perfect. However, as we show in the Appendix (Section 8.3), perfection requires the belief that rejections of $1/2$ by L types are more likely than rejections of $1/2$ by H types. That is, the equilibrium can be robust to small trembles only if higher probability is assigned to dominated rather than un-dominated actions. Under more plausible beliefs, convergence to the $\alpha_h = 1$ equilibrium is ruled out.

Proposition 3 is very relevant for a lab experiment and possibly for actual applications of mediation plans, where some lying seems inevitable. The proposition tells us that, when optimal mediation involves obfuscation, no peace probability in the neighborhood of the HMS equilibrium should be expected. It is important to stress that this only applies 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 whenever $q < (2\theta - 1)$ 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 constraint 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.[32]

[32] Without obfuscation, subjects learn their opponent's message from the mediator's recommendations. Hence, for example, we have that $\Delta_{Hh}(1/2) = (1/2 - \theta/2)\frac{\tau_H q}{\tau_H q + (1-\tau_L)(1-q)}\alpha_h + (1/2 - \theta)\frac{(1-\tau_L)(1-q)}{\tau_H q + (1-\tau_L)(1-q)}$. 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).

5.2 Sincerity and Peace: Data v/s Equilibrium Predictions

Having established the lack of predictive power of the HMS equilibrium, we ask if the data is better described by other equilibria. Figure 7 below superimposes the data, aggregated by session, to the equilibria in Figure 6. The data are represented by red spheres.

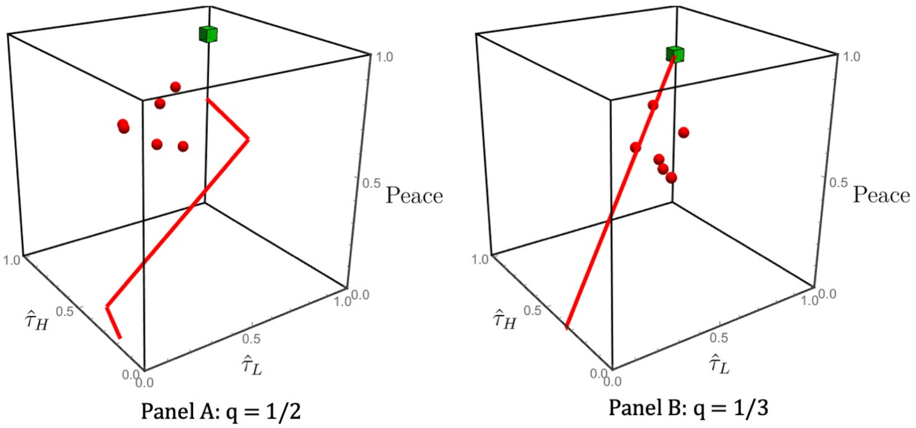


Figure 7: CM: Data and equilibria.

As we already know, in both parameterizations and all sessions, sincerity, by either type, and peace, all 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.[33] But note that peace would have occurred only if the opponent sent message l[34] 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

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

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

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 plausible that other 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.

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 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.[35]

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 the no communication (NC) treatment. In NC, 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 supplemental material (Section 10.3.4)). 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 time; 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).[36] 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.

[35] 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, if $u()$ is concave, 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.

[36] Under $q = 1/3$, we have only found equilibria in which the human mediator walks about 100 percent of the time following (h, h). The online Appendix, Section 9.2, characterizes equlibria of the HM game.

A second possible explanation for the behavior we observe is that the actions chosen by the subjects come at little individual cost. With the theory predicting that an H will reject any offer of 50 with probability 1, any noise results in more acceptances and more peace, and 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.

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.[37] 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. 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 8 and 9 are contour plots reproducing the loss from not best responding, as a percentage of the maximum possible payoff. Figure 8 refers to $q = 1/2$ and Figure 9 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 axes 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. 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.

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

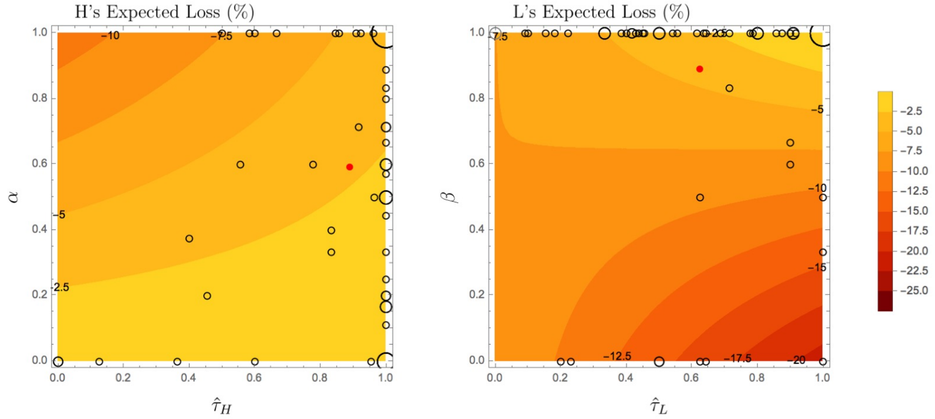


Figure 8: Losses relative to best responding; $q = 1/2$.

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 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 acceptance of the mediator's recommendations are 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.

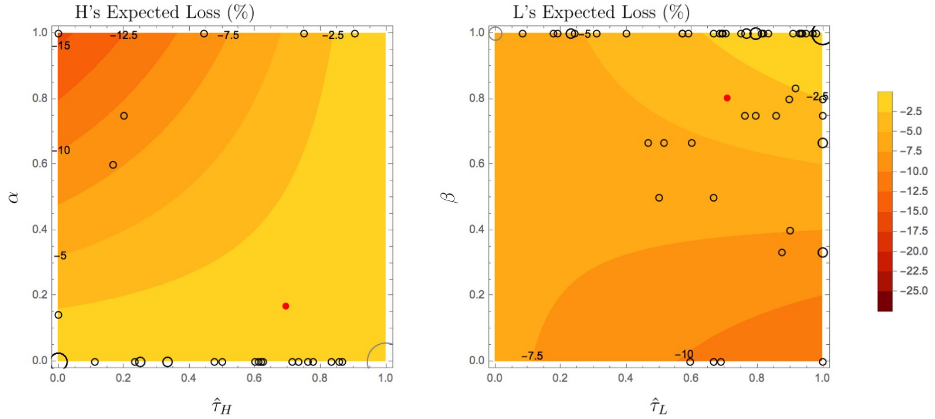


Figure 9: Losses relative to best responding; $q = 1/3$

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 because subjects choose strategies that limit their losses? How badly would subjects fare if they 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 distributions of observed messages, and the distributions of random acceptance decisions to the distributions of observed acceptances.[38] All eight resulting tests strongly reject the hypothesis that subjects' choices were random ($p < 0.001$ in all cases).

Figure 10 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 both cases; the equivalent numbers for

[38] At the individual observation level, both truthfulness and acceptances are coded as binary variables–either 0 or 1. Given the finite number of rounds, the observed average truthfulness and acceptance by subjects (conditioning on type) are discrete variables. To replicate this discreteness, we construct a random dataset (of the same size as the original) by drawing a sample of binary variables equally likely to be 0 or 1. We then compute the corresponding implied average truthfulness and acceptance rates for each subject, generating an "as if random" distribution, which we compare to the empirical 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 (random) population.

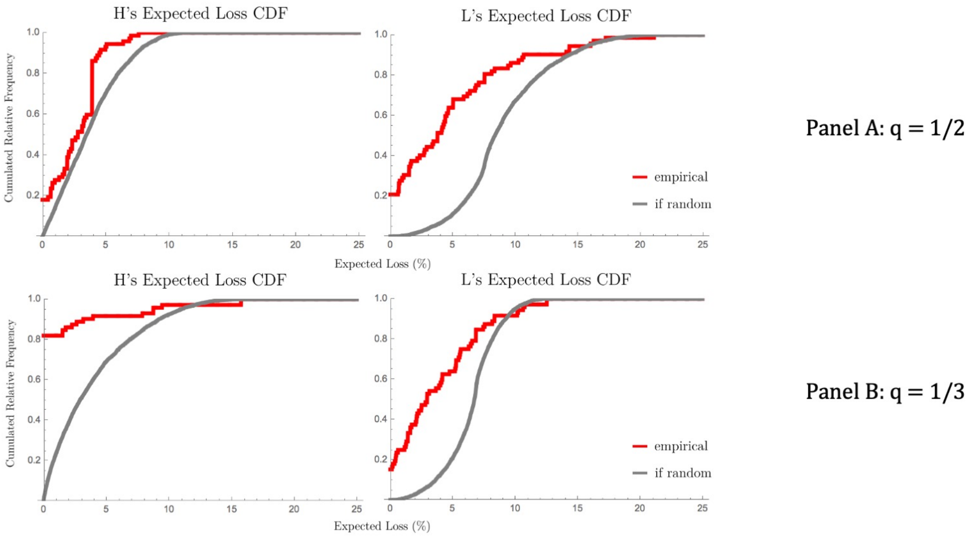


Figure 10: CDF's of losses, given observed play by others.

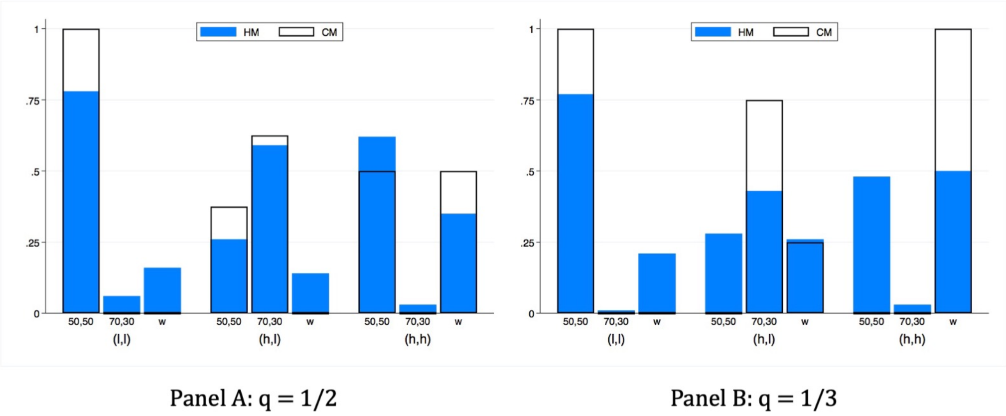
H players are 70 percent with $q = 1/2$ (v/s 93 percent in the data) and 69 percent with $q = 1/3$ (v/s 92 percent in the data). Experimental subjects are playing strategies that although not best responses are not far from them, in payoff space.

6 The HM Treatment

The focus of the experiment was the CM treatment, but we are also interested in the more exploratory HM treatment, where commitment was neither automatically present nor could develop through reputation or punishment. The outcomes are reflected in Figures 2 and 3 and in Tables 1 and 2: for L types, truthfulness under HM is intermediate between UC and CM, while the frequency of peace is somewhat lower than in the other two treatments. The problem for the subjects was very difficult, but looking at the data in more detail shows that interesting regularities did arise.

Figure 11 and Table 4 summarize the data. In both parameterizations, less than 7% of messages were silent, and for ease of reading we do not include them below (the full data set is reported in Table 10 in the supplemental material).[39] Figure 11 compares the mediation programs under HM and

[39] We have verified that the regularities we describe in the text are robust to plausible interpretations of silence (if silence is interpreted according to the prior, or in the same fashion regardless of q, or in line with equilibrium strategies).



Panel A: q = 1/2

Panel B: q = 1/3

Figure 11: HM v/s CM: Recommendations.

CM. The possible message pairs are aligned on the horizontal axis. For each pair, the figure shows the frequency of different recommendations seen in the data under HM (the blue columns) and in the optimal program under CM (the black profiles). Table 4 reports the players' strategies in the two mediation treatments, as well as the realized peace frequencies.

$q = 1/2$

	τ_H	τ_L	α_h^{50}	α_l^{50}	α_l^{30}	β_h^{50}	β_l^{50}	β_l^{30}	$Peace$
HM	0.80	0.38	0.68	0.44	0.41	0.87	0.94	0.89	0.50
CM	0.87	0.57	0.63	0.17	0.00	0.99	1.00	0.92	0.55

$q = 1/3$

	τ_H	τ_L	α_h^{50}	α_l^{50}	α_l^{30}	β_l^{50}	β_h^{50}	β_l^{30}	$Peace$
HM	0.65	0.49	0.56	0.55	0.38	0.92	0.81	0.70	0.40
CM	0.66	0.63	–	0.19	0.04	0.99	–	0.82	0.46

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

As in the CM program, HM's refusals to mediate (w) are more frequent after messages (h, h) for both parameterizations, and, less intuitively, at lower q for all message pairs. Such refusals are not consistently higher than in CM, minimizing potential concerns about distortions due to risk aversion. In fact, the clearest deviation is a bias towards peace after (h, h) messages, especially under $q = 1/3$ (when CM always offers w).

Players' strategies are close across the two treatments, for both values of q, but for two systematic differences: in HM, L is less sincere (τ_L is smaller), and H is more willing to accept HM's offer, especially after message l (α_l^{50} and α_l^{30} are higher). As in CM, peace is higher under $q = 1/2$.

The positive value of α_l^{30}, a dominated action, can only be imputed to noise, but occurrences of offers of 30 to H types are very few.[40] As for the other differences, a natural question is whether they are predicted by equilibrium behavior, in the absence of commitment by the mediator. In the online Appendix (Section 9.2) we characterize equilibria of the HM game that approximate what we observe in the data. Such equilibria align reasonably well with what we observe under $q = 1/2$ and correctly predict its higher frequency of peace, but cannot explain the mediator's choices under $q = 1/3$.[41] Behavior in the lab is more similar across the two parameterizations than equilibrium predicts. We conjecture that the regularities we see are better explained as reactions to the strategic uncertainty of a particularly challenging treatment.

When offered 50, H types are more likely to accept, a plausible reaction to strategic uncertainty since accepting reduces the probability of receiving the lowest payoff for any belief. Seemingly anticipating this, the human mediator walks out less than in the optimal CM program following h messages, and the decision to walk out is less sensitive to the content of messages. The joint behavior of H types and the human mediator reduces the L type's cost of lying. By messaging h, the L type is less likely to trigger walking out, and if 50 is offered, the opponent is more likely to accept, supporting more lying by the L type than in CM.

Predictably, but interestingly, the weaker punishment for lying translates into higher conflict: h messages are now more frequent, and even with the more indulgent human mediator induce more refusals to mediate than l messages. And it is L types that end up suffering most: under CM, $L - L$ pairs in the lab avoid conflict 75 percent of the times when $q = 1/2$ and 90 percent of the times when $q = 1/3$, but the fractions fall to 53 and 68 percent respectively under HM.[42]

HM's willingness to mediate after (h, h) and L types' incentive to lie are best responses to each other and to H types' acceptance of 50. H's behavior is suboptimal, but comes at very small individual cost. Figures 12 and 13 illustrate this point. The figures are parallel to Figures 8 and 9, now drawn for HM data–contour plots representing percentage losses from best responding, given the empirical strategies observed in the lab.[43] The square plots refer to H and L types, and map the frequency of

[40] Only about 10 percent of all offers to H types for both parameterizations are 30, and since α_l^{30} is calculated conditional on such offers, its importance is minor.

[41] In particular, they cannot rationalize the mediator's high acceptances to mediate after messages (h, h) under $q = 1/3$.

[42] While, in line with the observations above, $H - H$ types enjoy much higher peace under HM.

[43] In line with the rest of this section, the figures ignore the rare instances of silence, but we have verified that the

sincerity, on the horizontal axis, and of acceptances (of 50 by H's and of 30 by L's) on the vertical axis; the triangular plot is a simplex representing the mediator's recommendation frequencies in response to (h, h) messages: always w, at the apex; always $(70, 30)$ at the bottom left corner; always $(50, 50)$ at the bottom right corner.

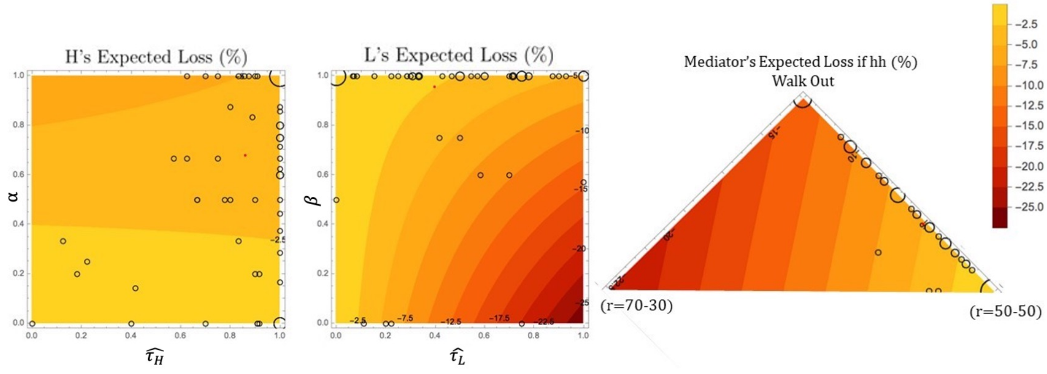


Figure 12: HM: Losses relative to best responding; $q = 1/2$

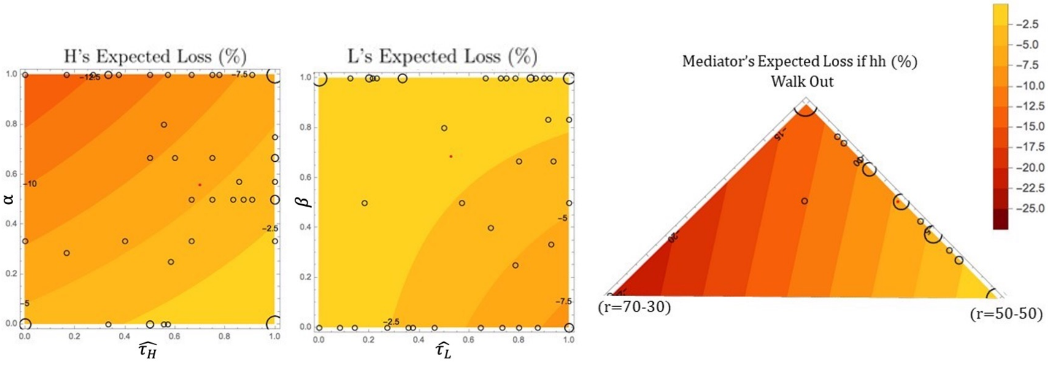


Figure 13: HM: Losses relative to best responding; $q = 1/3$

Under both parameterizations, L types' best response in the lab is to lie and message h and to accept all offers, and the mediator's best response, when facing messages (h, h), is to propose $(50, 50)$. For H types, sincerity and rejection of 50, rather than acceptance, are best responses, but as long as the message is sincere, the foregone gain from accepting 50 is small; the higher cost from acceptances

results extend to their inclusion. To avoid clutter, we do not reproduce here the plots describing the mediator's best response to messages (l, l) and (h, l)–$(50, 50)$ in both cases.

when $q = 1/3$ reflects the higher proportion of L types.

Summarizing, the most interesting lesson is the chain of incentives that leads from H's relatively high acceptances of 50 to the human mediator's frequent willingness to mediate when receiving messages (h, h), and thus to L types' incentive to lie. The final result is less peace than under CM and, especially, a different type of peace: the more permissive behavior ends up penalizing disproportionately the L types. While the high frequency of mediation seems to reflect the mediator's lack of commitment, our preferred reading is that it is in fact H's behavior that sustains the whole chain.[44]

7 Conclusions

The experiment tests the potential of a sophisticated mediation algorithm 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 communicate directly. We implement the optimal mediation algorithm 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, i.e. when the mediator is able to obfuscate the message received by the opposite party. 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, we find deviations from equilibrium, both when the optimal mediation program involves obfuscation and when it does not. Given the difficult game, the observation that the subjects' behavior has noise is not very surprising. Rather, what is interesting is that behavior is far from erratic, but deviations from best responses that cause participants only small payoff losses have significant

[44] In the equilibria we characterize, for $q = 1/3$, HM always refuses to mediate in response to (h, h). The behavior we see in the lab cannot be imputed solely to the lack of mediator's commitment.

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. Modifying the mediation program so that the best equilibrium is strict would not solve the multiple equilibrium problem, but would it improve outcomes? It would be interesting to think 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. It would be interesting to design 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 test their performance relative to the theoretically optimal mechanism with fully rational agents.

We complement our study of the optimal mediation mechanism, implemented by a computer program, with a treatment where mediation is provided by one of the subjects, in a design where commitment is not possible. We find a bias towards H types accepting an equal division of resources, and such a bias ripples into fewer refusals to mediate on the part of the human mediator, and less reason to be truthful for the L types. The final effect is a higher frequency of conflict, hitting primarily $L-L$ pairs, in clear contrast to optimal mediation.

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

Aumann, R. and S. Hart, 2003, "Long Cheap Talk", *Econometrica*, 71, 1619–1660.

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.

Barnett, J. and P. Treleavan, 2018, "Algorithmic Dispute Resolution–The Automation of Professional Dispute Resolution Using AI and Blockchain Technologies", *The Computer Journal*, 61, 399–408.

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., O. Board and K. Kawamura, 2007, Noisy Talk, Theoretical Economics 2, 395-440.

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.

Brams, S. and A. Taylor, 1996, "A Procedure for Divorce Settlements", *Mediation Quarterly*, 13, 191–205.

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.

Chassang S. and G. Padró I Miquel, 2019, Crime, Intimidation, and Whistleblowing: A Theory of Inference from Unverifiable Reports, The Review of Economic Studies, 86 2530–2553.

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.

Galanter, M., 2004, "The Vanishing Trial: An Examination of Trials and Related Matters in Federal and State Courts", *Journal of Empirical Legal Studies*, 1, 459–570.

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.

Ivanov, M., 2010, "Communication via a strategic mediator", *Journal of Economic Theory*, 145, 869–884.

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.

Lodder, A. and J. Zeleznikow, 2010, *Enhanced Dispute Resolution Through the Use of Information Technology*, Cambridge, UK: Cambridge University Press.

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., 1982, *Game Theory: Analysis of Conflict*, Cambridge, Ma and London, UK: Harvard University Press.

Myerson, R., 1991, "Optimal coordination mechanisms in generalized principal–agent problems", *Journal of Mathematical Economics*, 10, 67-81.

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.

Rauchhaus, R., 2006, "Asymmetric Information, Mediation, and Conflict Management", *World Politics*, 58, 207-241.

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.

Smith, A. and A. Stam, 2003, "Mediation and Peacekeeping in a Random Walk Model of Civil and Interstate War", *International Studies Review*, 5, 115-135.

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 our experimental design, 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 first establish the following lemma, which characterizes all truthful equilibria. We allow for general HM payoffs: HM 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.

Hence, in searching for equilibria, we need only consider that, following (h, l), HM offers $(\theta, 1-\theta)$ either w.p. 1 or w.p. 0. Allowing for $W > 0$ or $W = 0$ results in four cases total to consider. We show, in the supplemental material (Section 10.1), that the only equilibria are those stated in the lemma. □

Proof of Proposition 2. Equilibria 1 of Lemma A1, which only exist for $q \geq (2\theta - 1)$, involve peace $P_1 = 1 - q^2$. Equilibria 2 of Lemma A1, 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 in undominated strategies. In addition, 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 $(\tau_L = 1, \tau_H = 0)$, or, alternatively, we select equilibria with an arbitrarily small but positive probability of silence. Table 5 gives the full set of equilibria.

We describe here in detail the derivation of the equilibria for $q = 1/3$. The $q = 1/2$ case is discussed in the online Appendix (Section 9.1). 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.

	$q = 1/2$		$q = 1/3$
(i)	$\alpha_h = 1, \beta = 1, \hat{\tau}_L = 1, \hat{\tau}_H = 1$	(i)	$\beta = 1, \hat{\tau}_L = 1, \hat{\tau}_H = 1$
(ii)	$\alpha_h = 0, \beta = 1, \hat{\tau}_L = 1, \hat{\tau}_H = 1$	(ii)	$\alpha_l = 0, \beta = 1, \hat{\tau}_L \in (0, 1),$
(iii)	$\alpha_l = 0, \alpha_h = 0, \hat{\tau}_L = 0, \hat{\tau}_H \in [1/6, 4/15]$		$\hat{\tau}_H = 1/3 + (2/3)\hat{\tau}_L$
(iv)	$\alpha_l = 0, \alpha_h = 0, \beta = 1, \hat{\tau}_L \in (0, 1),$	(iii)	$\alpha_l = 0, \beta = 4/7, \hat{\tau}_L \in (0, 1),$
	$\hat{\tau}_H = 4/15 + (6/15)\hat{\tau}_L$		$\hat{\tau}_H = 1/3 - \hat{\tau}_L/3$
(v)	$\alpha_l = 0, \alpha_h = 0, \beta = 1, \hat{\tau}_L = 1, \hat{\tau}_H \in [2/3, 1)$	(iv)	$\alpha_l = 0, \hat{\tau}_L = 0, \hat{\tau}_H \leq 1/3$
(vi)	$\alpha_l = 0, \alpha_h = 0, \beta \in (0, 3/7), \hat{\tau}_L = 3/(18 - 35\beta),$		
	$\hat{\tau}_H = (1/6)(1 - 3/(18 - 35\beta))$		
(vii)	$\alpha_h = 0, \beta = 0, \hat{\tau}_L = 1/6, \hat{\tau}_H \leq 5/36$		

Table 5: CM: Equilibria in undominated strategies

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:[45]

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

In any equilibrium in which all accept 70, Ll player i will accept 30 with positive probability only if $EU_{Ll}(\textit{accept } 30) \geq EU_{Ll}(\textit{reject } 30)$ where $EU_{Ll}(\textit{accept } 30) = 30$ and $EU_{Ll}(\textit{reject } 30) = 35\Pr(j = L|h_j)$. Substituting (3):

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

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

[45] 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.

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

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

Hence:

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

As above, since we are considering the acceptance decision for 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 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 only L types would accept; 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, 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 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 given in the $q = 1/3$ column of Table 5. 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: $\Pr(j\ is\ L|\widehat{h}_j) = \frac{\Pr(\widehat{h}_j|j\ is\ L)\Pr(L)}{\Pr(\widehat{h}_j|j\ is\ L)\Pr(L)+\Pr(\widehat{h}_j|j\ is\ H)\Pr(H)} = \frac{[1-\tau_L-\sigma_L+(1/3)\sigma_L](2/3)}{[1-\tau_L=\sigma_L+(1/3)\sigma_L](2/3)+[\tau_H+(1/3)\sigma_H](1/3)} = \frac{2(1-\widehat{\tau}_L)}{2(1-\widehat{\tau}_L)+\widehat{\tau}_H}$. 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.

8.3 Trembling-hand Perfection

If $(2\theta - 1) < q < (2\theta - 1)/\theta$, the HMS equilibrium is trembling-hand perfect, but this requires beliefs about trembles that assign higher probability to dominated actions. 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. In the HMS equilibrium, the L type ex-post participation constraint is slack in equilibrium; the three incentive constraints that bind 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 (sending message l and then rejecting $1/2$). We write below the three conditions that must be satisfied for the prescribed strategies to be best responses, given trembles around equilibrium behavior. 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 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:

$$(1/2 - \theta/2)q\left[q_H\tau_H\alpha_h^{1/2} + q_M(1-\tau_H)\alpha_l^{1/2}\right] \geq (\theta - 1/2)(1-q)\left[q_H(1-\tau_L)\beta_h^{1/2} + q_M\tau_L\beta_l^{1/2}\right] \quad (8)$$

where, from (1) in the text: $q_M = (\frac{1-\theta}{2\theta-1})(\frac{1+q-2\theta}{\theta-q})$ and $q_H = (\frac{1-q}{q})(\frac{1+q-2\theta}{\theta-q})$. In addition, both types prefer to be truthful. For a player of type L we require $EU_L(l) \geq EU_L(h)$ where:

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

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

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

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

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

Consider trembles such that: $\alpha_h^{\theta} = 1 - a_h^{\theta}/n$, $\alpha_h^{1/2} = 1 - a_h^{1/2}/n$, $\alpha_l^{1/2} = a_l^{1/2}/n$, $\alpha_l^{1-\theta} = a_l^{1-\theta}/n$, $\beta_h^{\theta} = 1 - b_h^{\theta}/n$, $\beta_h^{1/2} = 1 - b_h^{1/2}/n$, $\beta_l^{1/2} = 1 - b_l^{1/2}/n$, $\beta_l^{1-\theta} = 1 - b_l^{1-\theta}/n$, $\tau_H = 1 - t_H/n$, $\tau_L = 1 - t_L/n$. We search for a vector of positive contants $\{t_H, t_L, a_h^{1/2}, a_h^{\theta}, a_l^{1/2}, a_l^{1-\theta}, b_l^{1/2}, b_l^{1-\theta}, b_h^{1/2}, b_h^{\theta}\}$ such that:

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

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

A vector that satisfies these conditions does exist. For example, at the experimental parameters of $q = 1/2$ and $\theta = 0.7$, all three conditions are satisfied at $\{t_H = 1, t_L = 1, a_h^{1/2} = 1, a_h^{\theta} = 1, a_l^{1/2} = 1, a_l^{1-\theta} = 1, b_l^{1/2} = 3, b_l^{1-\theta} = 1, b_h^{1/2} = 4, b_h^{\theta} = 1\}$. Note that beliefs assign higher probability to trembles that result in L types' rejections than to trembles that result in H types' rejections. This is not an anomaly; this is necessary for THP and does not depend on the experimental parameterization:

Proposition THP. *Suppose* $(2\theta-1) < q < (2\theta-1)/\theta$. *Then the HMS equilibrium can be trembling-hand perfect only if along the sequence of trembles* $\alpha_h^{1/2} > \min(\beta_l^{1/2}, \beta_h^{1/2})$, *or* $a_h^{1/2}/n < \max(b_h^{1/2}/n, b_l^{1/2}/n)$.

Proof. Condition (8) corresponds to $q_M[q_H\tau_H\alpha_h^{1/2} + q_M(1-\tau_H)\alpha_l^{1/2}] \geq q_H[q_H(1-\tau_L)\beta_h^{1/2} + q_M\tau_L\beta_l^{1/2}]$. Note that $q < (2\theta-1)/\theta$ implies $(1-\theta)/(2\theta-1) < (1-q)/q$, and thus $q_M < q_H$. In addition, $\alpha_l^{1/2}$ must converge to 0 in equilibrium. All agents' choices are binary choices, and thus all small enough trembles–all trembles that assign lower probability to the suboptimal action–must have probability lower then 1/2. Thus, along the sequence of trembles, $\alpha_l^{1/2} = a_l^{1/2}/n < 1/2 < \alpha_h^{1/2} = 1 - a_h^{1/2}/n$. A necessary condition for (8) is then $q_Mq_H[\alpha_h^{1/2}[\tau_H + (1-\tau_H)] > q_Mq_H[\beta_h^{1/2}(1-\tau_L) + \beta_l^{1/2}\tau_L)]$ or $\alpha_h^{1/2} > \min(\beta_l^{1/2}, \beta_h^{1/2})$, which is equivalent to $a_h^{1/2}/n < \max(b_h^{1/2}/n, b_l^{1/2}/n)$. □

The result in the proposition is problematic because accepting 1/2 is dominant for L, but not for Hh. Thus, a necessary condition for convergence to the HMS equilibrium is beliefs that assign higher probability to deviation from a dominant rather than a non-dominant action.

For Online Publication

9 Online Appendix

9.1 Multiple Equilibria under CM: The $q = 1/2$ Case

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 $q = 1/2$:

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

Ll accepts with positive probability if:

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

or:

$$6\tau_H > 1-\tau_L \Longrightarrow \beta = 1,\ 6\tau_H < 1-\tau_L \Longrightarrow \beta = 0, \text{ and } 6\tau_H = 1-\tau_L \Longrightarrow \beta \in [0,1]. \tag{9}$$

(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:

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

where:

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

and:

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

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

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

and

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

Hh will accept 50 with positive probability if:

$$15\Pr(j\ is\ H|(50,50),\ j \text{ accepts } 50, h_i) \geq 20\Pr(j\ is\ L|(50,50),\ j \text{ accepts } 50, h_i) \tag{10}$$

or:

$$\begin{aligned}
15(4\tau_H\alpha_h + 3(1-\tau_H)\alpha_l) = 20(4(1-\tau_L) + 3\tau_L) &\Longrightarrow \alpha_h \in [0,1] \\
15(4\tau_H\alpha_h + 3(1-\tau_H)\alpha_l) < 20(4(1-\tau_L) + 3\tau_L) &\Longrightarrow \alpha_h = 0, \\
15(4\tau_H\alpha_h + 3(1-\tau_H)\alpha_l) > 20(4(1-\tau_L) + 3\tau_L) &\Longrightarrow \alpha_h = 1
\end{aligned} \tag{11}$$

Condition (10) 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:

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

and will accept 50 with positive probability if:

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

or:

$$\begin{aligned} 15(3\tau_H\alpha_h + 8(1-\tau_H)\alpha_l) = 20(3(1-\tau_L) + 8\tau_L) &\implies \alpha_l \in [0,1] \\ 15(3\tau_H\alpha_h + 8(1-\tau_H)\alpha_l) < 20(3(1-\tau_L) + 8\tau_L) &\implies \alpha_l = 0 \\ 15(3\tau_H\alpha_h + 8(1-\tau_H)\alpha_l) > 20(3(1-\tau_L) + 8\tau_L) &\implies \alpha_l = 1 \end{aligned} \tag{12}$$

Conditions (9), (11), and (12) 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 (9). 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:

$$\begin{aligned} EU_H(h) &= (1/2)[\tau_H 35 + (1-\tau_H)((5/8)35 + (3/8)(50\alpha_l + 35(1-\alpha_l)))] + (1/2)70 \\ EU_H(l) &= (1/2)[\tau_H 35 + (1-\tau_H)(50\alpha_l^2 + 35(1-\alpha_l^2))] + \\ &\quad (1/2)[\tau_L(\alpha_l 50 + (1-\alpha_l)70) + (1-\tau_L)((5/8)70 + (3/8)(50\alpha_l + 70(1-\alpha_l)))] \\ EU_L(l) &= (1/2)[\tau_H(5/8)30\beta + (1-\tau_H)50\alpha_l] + (1/2)[\tau_L 50 + (1-\tau_L)((5/8)(30\beta + 35(1-\beta)) + (3/8)50)] \\ EU_L(h) &= (1/2)[(1-\tau_H)((3/8)50\alpha_l] + \\ &\quad (1/2)[\tau_L((5/8)(70\beta + 35(1-\beta)) + (3/8)50) + (1-\tau_L)(50/2 + 35/2)] \end{aligned} \tag{13}$$

As before, four conditions, (12), (9), 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 (12):

$$6(1-\tau_H)\alpha_l = 3 + 5\tau_L \Longrightarrow \tau_H = 1 - \left(\frac{3+5\tau_L}{6\alpha_l}\right). \tag{14}$$

Substituting (14) in (13), 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 (14) is violated. Thus $\alpha_l \in (0,1)$ is impossible.[46]

(2) Suppose then $\alpha_l = 1$. From (12), it follows that:

$$\tau_H \leq (1/2) - (5/6)\tau_L. \tag{15}$$

Note that there cannot be an equilibrium with $\alpha_l = 1$ if L prefers sincerity and thus $\tau_L = 1$. From (13):

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

an expression that is minimal when τ_H is maximal. By (15), 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 (15) 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

[46] 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 (12) an H player who lied would still reject 50.

(13) and (9), we can identify the full set of equilibria:[47]

(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.

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.[48] 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

[47] 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.

[48] 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 (9) can be rewritten as before.

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.

9.2 Human Mediation: Equilibria (at the Experimental Parameter Values)

We concentrate on equilibria for which: $\tau_H = 1$, $r(l,l) = (50,50)$ w.p.1, and dominant acceptance strategies are taken (both players accept 70, L accepts 50, and H rejects 30). We prove the following proposition, which gives the equilibria discussed in Section 6.

Proposition HM. (i) $\underline{q = 1/2.}$ *There is an equilibrium in which* $r(h,h) = w$ *w.p. 1;* $\tau_L = 1$*;* α_h *sufficiently low,* $\alpha_l = 0$*,* $\beta_l = 1$*,* β_h *arbitrary;* $P = \frac{3}{4}$*.* $\underline{q = 1/3.}$ *There is an equilibrium in which* $r(h,h) = w$ *w.p. 1;* $\tau_L = \frac{2}{3}$*;* α_h *sufficiently low,* $\alpha_l = 0$*,* $\beta_l = 1$*,* β_h *sufficiently low;* $P = \frac{56}{81} \approx 0.691$*.* (ii) $\underline{q = 1/2.}$ *There is an equilibrium in which* $r(h,h) = w$ *w.p.* $p_w \approx 0.535$*,* $r(h,h) = (50,50)$ *w.p.* $1 - p_w$*;* $\tau_L \approx 0.565$*;* $\alpha_h \approx 0.580$*,* $\alpha_l = 0$*,* $\beta_l = 1$*,* β_h *arbitrary;* $P \approx 0.605$*.*

To prove the proposition, we make use of Lemma A3 below. The significance of this lemma is that any PBE with the features of interest in which $r(h,l) = (70,30)$ w.p. > 0 is such that (1) $r(l,l) = (50,50)$ w.p. 1, which is accepted w.p. 1, (2) $r(h,l) = (70,30)$ w.p. 1., which is accepted w.p. 1, (3) $\alpha_l = 0$, (4) $\beta_l = 1$, and (5) $\beta_h = 1$ if $r(h,h) = (70,30)$ w.p. > 0 (otherwise any β_h is optimal for L). Hence, in addition to confirming the optimality of $\tau_H = 1$ and finding an optimal τ_L, it remains only to determine (a) HM's strategy following (h,h), and (b) α_h, H's probability of accepting 50 after messaging h. Furthermore, the optimality conditions for (a) and (b) are given by part (v) of the lemma.

Lemma A3. *In any PBE such that* $\tau_H = 1$*,* $r(l,l) = (50,50)$ *w.p. 1, and dominant acceptance strategies are taken, the following must hold: (i) if* $r(h,l) = (70,30)$ *w.p.* > 0*, then* $\beta_l = 1$ *and thus* $r(h,l) = (70,30)$ *is accepted w.p. 1; (ii) if* $r(h,h) = (70,30)$ *w.p.* > 0 *(randomizing which player receives 30), then* $\beta_h = 1$*; (iii) if* $r(h,l) = (70,30)$ *w.p.* > 0*, then* $r(h,l) = (70,30)$ *w.p. 1; (iv) if* $r(h,l) = (70,30)$ *w.p.* > 0*, then* $\alpha_l = 0$*; and (v):*

- *HM prefers* $r(h,h) = (70,30)$ *to* $r(h,h) = w$ *if and only if*

$$\begin{aligned}(1-q_h)\beta_h \geq \frac{1}{2} \iff& \\ \tau_L \leq \frac{1-2q}{1-q} &\text{ if } \beta_h = 1,\end{aligned}$$

  where $q_h = \frac{q}{q+(1-\tau_L)(1-q)} \in [q,1]$.

- *HM prefers* $r(h,h) = (50,50)$ *to* $r(h,h) = w$ *if and only if*

$$q_h^2\alpha_h^2 + 2q_h(1-q_h)\alpha_h + (1-q_h)^2 \geq \frac{1}{2},$$

  where $q_h = \frac{q}{q+(1-\tau_L)(1-q)} \in [q,1]$.

- *HM prefers* $r(h,h) = (50,50)$ *to* $r(h,h) = (70,30)$ *if and only if*

$$q_h^2\alpha_h^2 + 2q_h(1-q_h)\alpha_h + (1-q_h)^2 \geq (1-q_h)\beta_h,$$

  where $q_h = \frac{q}{q+(1-\tau_L)(1-q)} \in [q,1]$.

- *If* $r(h,l) = (70,30)$ *w.p.* $>$ *0 and* $r(h,h) = (50,50)$ *w.p.* $>$ *0, then* H *prefers to accept 50 after messaging* h *if and only if*

$$\tau_L \geq 1 - \frac{3}{4}(\frac{q}{1-q})\alpha_h.$$

Proof. *(i)* Define $\Delta_{Ll}(30)$ as the expected utility difference between accepting and rejecting 30 for L who messages l. Assuming $\tau_H = 1$, that dominant acceptance strategies are taken, and $r(h,l) = (70,30)$ w.p. > 0 (which implies Bayes rule applies),

$$\Delta_{Ll}(30) \propto p30_{hl}(q30 - (1-q)(1-\tau_L)5) - \tilde{p}30_{ll}(1-q)\tau_L 5,$$

where $p30_{hl}$ is the probability with which $r(h,l) = (70,30)$ and $\tilde{p}30_{ll}$ is *one-half* the probability with which $r(l,l) = (70,30)$ (i.e. the probability $(70,30)$ is offered *and* it is this player who is offered 30). Thus, if $\tilde{p}30_{ll} = 0$ and $p30_{hl} > 0$, $\Delta_{Ll}(30)$.

(ii) If $r(h,h) = (70,30)$ w.p. > 0, then players who message h and receive 30 know the other player messaged h. Since we assume $\tau_H = 1$, the belief that the other player is H is given by

$q_h = \frac{q}{q+(1-\tau_L)(1-q)}$, which is between q and 1. Since the other player will always accept 70, the expected payoff of accepting is 30. The expected payoff of rejecting is $q_h 0 + (1-q_h)35$, which is strictly less than 30.

(iii) Since $r(h,l) = (70,30)$ is accepted w.p. 1 by part *(i)*, it cannot be that $r(h,l) = w$ w.p. > 0 since it yields a lower payoff for HM.

We suppose that $r(h,l) = (50,50)$ w.p. > 0 in equilibrium, and derive a contradiction. Since $r(h,l) = (70,30)$ is accepted w.p. 1, $r(h,l) = (50,50)$ must be accepted w.p. 1 (or else HM would never offer it). Hence, if $r(h,h) = (50,50)$ w.p. > 0, it is accepted w.p. 1 as it corresponds to the same information set for those who message h.

Case 1: Suppose $r(h,h) = (50,50)$ w.p. 0. This cannot be an equilibrium since $r(h,h) = (50,50)$ yields a strictly higher payoff for HM than $r(h,h) = w$ and $r(h,h) = (70,30)$ (H type will always reject 30).

Case 2: Suppose $r(h,h) = (50,50)$ w.p. > 0. Since this yields a higher payoff for HM than $r(h,h) = w$ and $r(h,h) = (70,30)$, it must be that $r(h,h) = (50,50)$ w.p. 1. But this cannot be an equilibrium because then the L-type would strictly prefer to message h (all offers are accepted and types messaging h would sometimes get more), and then the H type would prefer to reject an offer of 50.

(iv) Assume $\tau_H = 1$ and $r(h,l) = (70,30)$ w.p. > 0, which implies $r(h,l) = (70,30)$ w.p. 1 by part *(iii)*. Hence, if H deviates and messages l and receives 50, she knows her opponent is L and hence it is strictly optimal to reject, i.e. $\alpha_l = 0$.

(v) For simplicity, subtract 20 from HM's payoffs so that rejection yields 0, walking out yields 20, and acceptance yields 40. Let uW_{hh}, $u50_{hh}$, and $u70_{hh}$ be HM's expected payoffs from walking out, offering $(50,50)$, and offering $(70,30)$, respectively, after observing (h,h). These are given as

$$uW_{hh} = 20,$$

$$u50_{hh} = 40[q_h^2 \alpha_h^2 + 2q_h(1-q_h)\alpha_h + (1-q_h)^2],$$

$$u70_{hh} = 40[(1-q_h)\beta_h].$$

HM prefers $r(h,h) = (70,30)$ to $r(h,h) = w$ if and only if

$$\begin{aligned} 40(1-q_h)\beta_h \geq 20 &\iff \\ (1-q_h)\beta_h \geq \frac{1}{2}. \end{aligned}$$

If $\beta_h = 1$, then this condition becomes

$$\begin{aligned} (1-q_h)\beta_h \geq \frac{1}{2} &\iff \\ \frac{(1-\tau_L)(1-q)}{q+(1-\tau_L)(1-q)} \geq \frac{1}{2} &\iff \\ \frac{1-2q}{1-q} \geq \tau_L \end{aligned}$$

HM's other indifference conditions are immediate.

Define $\Delta_{Hh}(50)$ as the expected utility difference between accepting and rejecting 50 for H who messages h. Assuming $\tau_H = 1$, that dominant acceptance strategies are taken, $r(h,l) = (70,30)$ w.p. > 0 (which implies $r(h,l) = (50,50)$ w.p. 0 by part *(iii)*), and $r(h,h) = (50,50)$ w.p. > 0, then

$$\begin{aligned} \Delta_{Hh}(50) \propto & p50_{hh}\{q\tau_H\alpha_h 15 - (1-q)(1-\tau_L)20\} + \\ & p50_{hl}\{q(1-\tau_H)\alpha_l 15 - (1-q)\tau_L 20\} \end{aligned}$$

and thus

$$\begin{aligned} \Delta_{Hh}(50) \geq 0 &\iff \\ q\tau_H\alpha_h 15 - (1-q)(1-\tau_L)20 \geq 0 &\iff \\ \tau_L \geq 1 - \frac{3}{4}(\frac{q}{1-q})\alpha_h. \end{aligned}$$

□

Proof of Proposition HM. *(i)* We search for equilibria in undominated strategies in which $\tau_H = 1$, $r(l,l) = (50,50)$ w.p. 1, $r(h,l) = (70,30)$ w.p. > 0, and $r(h,h) = w$ w.p. 1.

In such an equilibrium, by Lemma A3, it must be that $r(h,l) = (70,30)$ w.p. 1 and $\beta_l = 1$.

Assuming all of these conditions, the expected payoffs to L of messaging l and h, respectively, are:

$$\begin{aligned} EU_L(l) &= q\{\tau_H 30 + (1-\tau_H)(\alpha_l 50 + (1-\alpha_l)0)\} + (1-q)\{\tau_L 50 + (1-\tau_L)30\} \\ &= 30 + \tau_L 20 - q\tau_L 20 \end{aligned}$$

and

$$\begin{aligned} EU_L(h) &= q\{\tau_H 0 + (1-\tau_H)0\} + (1-q)\{\tau_L 70 + (1-\tau_L)(35)\} \\ &= 35 + \tau_L 35 - q35 - q\tau_L 35, \end{aligned}$$

and so

$$\begin{gathered} EU_L(l) \geq EU_L(h) \iff \\ 30 + \tau_L 20 - q\tau_L 20 \geq 35 + \tau_L 35 - q35 - q\tau_L 35 \iff \\ \frac{q35-5}{15-q15} \geq \tau_L. \end{gathered}$$

Hence, if $q = \frac{1}{2}$, this is always satisfied strictly, meaning it must be that $\tau_L = 1$.

If $q = \frac{1}{3}$, then the condition becomes $EU_L(l) \geq EU_L(h) \iff \frac{2}{3} \geq \tau_L$. If $\frac{2}{3} > \tau_L$, then $EU_L(l) > EU_L(h)$ and it must be that $\tau_L = 1$, a contradiction. If $\frac{2}{3} < \tau_L$, then $EU_L(l) < EU_L(h)$ and it must be that $\tau_L = 0$, a contradiction. Hence it must be that $\tau_L = \frac{2}{3}$.

Following (h, h), HM must prefer to walk out rather than offer $(70, 30)$:

$$(1-q_h)\beta_h \leq \frac{1}{2}.$$

When $q = \frac{1}{2}$, $\tau_L = 1$ and $q_h = 1$, so this condition is satisfied for any β_h. When $q = \frac{1}{3}$, $\tau_L = \frac{2}{3}$ and $q_h = \frac{3}{7}$, so the condition is satisfied for sufficiently low β_h.

Following (h, h), HM must prefer to walk out rather than offer $(50, 50)$:

$$q_h^2\alpha_h^2 + 2q_h(1-q_h)\alpha_h + (1-q_h)^2 \leq \frac{1}{2},$$

which is satisfied for both values of q for sufficiently low α_h.

Finally, we must now check that $\tau_H = 1$ is optimal. First note that if H messages l and receives

50, it is optimal for her to reject, i.e. $\alpha_l = 0$, since she will be facing an L type and so she would receive 70. So H will always reject 50 and below and accept 70 and above. Hence, no matter the message she sends, she will always receive 35 against an H-type and 70 against an L-type. So H is indifferent between messages, meaning $\tau_H = 1$ is optimal in particular.

Summarizing, for $q = \frac{1}{2}$, there is an equilibrium involving $\tau_L = 1$, α_h sufficiently low, $\alpha_l = 0$, $\beta_l = 1$, β_h arbitrary, and the resulting probability of peace is $1 - q^2 = \frac{3}{4}$. For $q = \frac{1}{3}$, there is an equilibrium involving $\tau_L = \frac{2}{3}$, α_h sufficiently low, $\alpha_l = 0$, $\beta_l = 1$, β_h sufficiently low, and the resulting probability of peace is

$$\begin{aligned}1 - (q^2 + 2q(1-q)(1-\tau_L) + (1-q)^2(1-\tau_L)^2) = \\ 1 - (\frac{1}{9} + 2(\frac{1}{3})(\frac{2}{3})\frac{1}{3} + (\frac{2}{3})^2(\frac{1}{3})^2) = \\ \frac{56}{81} \approx 0.691.\end{aligned}$$

(ii) We search for equilibria in undominated strategies in which $\tau_H = 1$, $r(l,l) = (50,50)$ w.p. 1, $r(h,l) = (70,30)$ w.p. > 0, and $r(h,h) = w$ w.p. $p_w \in (0,1)$ and $r(h,h) = (50,50)$ w.p. $1 - p_w$.

By Lemma A3, since $r(h,l) = (70,30)$ w.p. > 0 and $r(h,h) = (50,50)$ w.p. > 0, $r(h,l) = (70,30)$ w.p.1 and H prefers to accept 50 after messaging h if and only if

$$\tau_L \geq 1 - \frac{3}{4}(\frac{q}{1-q})\alpha_h.$$

Following (h,h), HM must be indifferent between walking out and offering $\{50,50\}$:

$$q_h^2\alpha_h^2 + 2q_h(1-q_h)\alpha_h + (1-q_h)^2 = \frac{1}{2}.$$

If $\alpha_h = 0$, then this cannot be satisfied (the LHS is $(1-q_h)^2$, and $q_h \in [q,1]$). If $\alpha_h = 1$, then this can also not be satisfied as the LHS equals $q_h^2 + 2q_h(1-q_h) + (1-q_h)^2 = q_h^2 + 2q_h - 2q_h^2 + 1 - 2q_h + q_h^2 = 1$. Hence, a necessary condition is that

$$\tau_L = 1 - \frac{3}{4}(\frac{q}{1-q})\alpha_h.$$

Plugging this into HM's indifference condition yields a non-linear equation in α_h where the LHS is strictly increasing in α_h. Solving the equation numerically gives a unique solution for each parameter value:

- $q = \frac{1}{2}$: $\alpha_h^* \approx 0.580$ and $\tau_L^* \approx 0.565$
- $q = \frac{1}{3}$: $\alpha_h^* \approx 0.580$ and $\tau_L^* \approx 0.783$

Since $\tau_L \in (0,1)$, we must have that $EU_L(l) = EU_L(h)$, where

$$\begin{aligned} EU_L(l) &= q\{\tau_H 30 + (1-\tau_H)(\alpha_l 50 + (1-\alpha_l)0)\} + (1-q)\{\tau_L^* 50 + (1-\tau_L^*)30\} \\ &= 30 + \tau_L^* 20 - q\tau_L^* 20 \end{aligned}$$

and

$$\begin{aligned} EU_L(h) =& q\{\tau_H(p_w 0 + (1-p_w)50\alpha_h^*) + (1-\tau_H)0\} \\ &+ (1-q)\{\tau_L^* 70 + (1-\tau_L^*)(p_w 35 + (1-p_w)50)\} \\ =& q(1-p_w)50\alpha_h^* + (1-q)\{\tau_L^* 70 + 50 - 15p_w - \tau_L^* 50 + \tau_L^* p_w 15\}. \end{aligned}$$

Equating $EU_L(l) = EU_L(h)$ gives an equation in p_w where the RHS is strictly decreasing in p_w. For $q = \frac{1}{3}$, and the values of α_h^* and τ_L^* we found, there is no solution in p_w, and thus no equilibrium. For $q = \frac{1}{2}$, there is a solution, $p_w \approx 0.535$.

We must check that HM prefers walking out to offering $(70, 30)$:

$$(1-q_h)\beta_h \leq \frac{1}{2}.$$

For the value of τ_L^* we found, $q_h \approx 0.697$, and so this is satisfied for any β_h.

Finally, we must now check that $\tau_H = 1$ is optimal. First note that if H messages l and receives 50, it is optimal for her to reject, i.e. $\alpha_l = 0$, since she will be facing an L type and so she would receive 70. So if she messages l, she will receive 70 against an L who messages l, 70 against an L who messages h (she would be offered 30, which she will reject), and 35 against an H (H messages h, so she will be offered 30 which she will reject). So messaging l gives H a payoff of $EU_H(l) = (1-q)70 + q35 = 70 - 35q$. If instead, she messages h, then against another H, either HM will walk out in which case she gets 35 or she will be offered 50, which she is indifferent between accepting and rejecting; suppose she rejects and receives 35. Against an L who messages l, she will be offered 70 which she will accept and receive (the L will accept 30 ($\beta_l = 1$) by Lemma A3. Against an L who messages h, either HM will walk out in which case she gets 70 or she will be offered 50, which she is indifferent between accepting and rejecting; suppose she rejects and receives 70. Hence, messaging h gives H a payoff of $EU_H(h) = (1-q)70 + q35 = 70 - 35q$. So H is indifferent between messages, meaning $\tau_H = 1$ is optimal in particular.

Summarizing, for $q = \frac{1}{3}$, there is no equilibrium. For $q = \frac{1}{2}$, there is an equilibrium involving $\tau_L \approx 0.565$, $\alpha_h \approx 0.580$, $\alpha_l = 0$, $\beta_l = 1$, β_h arbitrary, $p_w \approx 0.535$, and the resulting probability of peace is

$$\begin{gathered} q^2(1-p_w){\alpha_h}^2 + \\ 2q(1-q)\{(1-\tau_L)(1-p_w)\alpha_h + \tau_L\} + \\ (1-q)^2\{(1-\tau_L)^2(1-p_w) + 2\tau_L(1-\tau_L) + \tau_L^2\} \approx 0.605. \end{gathered}$$

□

9.3 Unmediated Communication in the Lab: Equilibria

We focus on equilibria in undominated strategies where no player demands w, and, given $\theta/2 > 1-\theta$, H types never demand $1-\theta$. The logic guiding the characterization of the equilibria is straightforward. Whether different demand strategies are best responses to each other depends on the posterior probabilities of the opponent's types, given the messages. The necessary restrictions on the posterior probabilities amount to restrictions on the probabilities τ_T and σ_T. In equilibrium, messages are random, and both types are indifferent over sending any of the three messages. Denote by $\delta_d(T, m, m')$ the probability that type T who has sent message m and received message m' demands d. Then:

Proposition A1.

(1) *For any $q < (2\theta-1)/\theta$, $\theta/2 > 1-\theta$, there exist equilibria in undominated strategies such that, at the demand stage:*

$$\begin{aligned} &\delta_\theta(H, m, m') = 1 \textit{ for all } m, m' \\ &\delta_\theta(L, m, m') = 1 - \delta_{1-\theta}(L, l, m') = 2\left(1 - \frac{1-\theta}{\theta(1-\pi_m)}\right) \end{aligned}$$

where π_m is the posterior probability that a player who sent message m is of type H, or:

$$\pi_l = \frac{(1-\sigma_H-\tau_H)q}{(1-\sigma_H-\tau_H)q+\tau_L(1-q)}; \quad \pi_h = \frac{q\tau_H}{q\tau_H+(1-q)(1-\sigma_L-\tau_L)}; \quad \pi_s = \frac{q\sigma_H}{q\sigma_H+(1-q)\sigma_L}.$$

At the message stage, $(\tau_L+\sigma_L) \in (0,1)$, $\sigma_L > 0$, and for any such τ_L and σ_L, τ_H and σ_H satisfy the constraints

$$\begin{aligned}
&\tau_H \geq \max\left[\left(\frac{3\theta-2}{2(1-\theta)}\right)\left(\frac{1-q}{q}\right)(1-\sigma_L-\tau_L), 1-\sigma_H-\left(\frac{2\theta-1}{1-\theta}\right)\left(\frac{1-q}{q}\right)\tau_L\right] \qquad (16)\\
&\tau_H \leq \min\left[\left(\frac{2\theta-1}{1-\theta}\right)\left(\frac{1-q}{q}\right)(1-\sigma_L-\tau_L), 1-\sigma_H-\left(\frac{3\theta-2}{2(1-\theta)}\right)\left(\frac{1-q}{q}\right)\tau_L\right]\\
&\sigma_H \in \left[\left(\frac{3\theta-2}{2(1-\theta)}\right)\left(\frac{1-q}{q}\right)\sigma_L, \left(\frac{2\theta-1}{1-\theta}\right)\left(\frac{1-q}{q}\right)\sigma_L\right]\\
&(\tau_H+\sigma_H) \in (0,1).
\end{aligned}$$

Given θ and q, the ex ante probability of peace, P is constant and given by:

$$P = \frac{[\theta(5-q)-2][2-\theta(3-q)]}{\theta^2}.$$

(2) *If $q \leq 2\theta-1$, there exist equilibria in undominated strategies such that, at the demand stage:*

$$\delta_\theta(H,m,m') = 1 \textit{ for all } m, m'$$

$$\delta_{1/2}(L,m,m') = 1 \textit{ for all } m, m'$$

At the message stage, $(\tau_L+\sigma_L) \in (0,1)$, $\sigma_L > 0$, and for any such τ_L and σ_L, τ_H and σ_H satisfy the constraints

$$\begin{aligned}
&\tau_H \in \left[1-\sigma_H-\left(\frac{2\theta-1}{1-\theta}\right)\left(\frac{1-q}{q}\right)\tau_L, \left(\frac{2\theta-1}{1-\theta}\right)\left(\frac{1-q}{q}\right)(1-\sigma_L-\tau_L)\right]\\
&\sigma_H \leq \left(\frac{2\theta-1}{1-\theta}\right)\left(\frac{1-q}{q}\right)\sigma_L \qquad (17)\\
&(\tau_H+\sigma_H) \in (0,1).
\end{aligned}$$

The ex ante probability of peace is $P = (1-q)^2$.

Proof.

The logic of the proof is straightforward, but the derivation is cumbersome. We begin by proving result (1). It is convenient to start by ignoring the option of silence.

(1). Step 1. Suppose $m \in \{l, h\}$ only. Denote by $S_{T,m|m'}(x)$ the expected share of a player of type T who sent message m, received message m' and demands x, where $x \in X = \{1-\theta, 1/2, \theta\}$, the set of possible (undominated) demands. Ignoring silence, there are eight different $(T, m|m')$ combinations, which we distinguish by labels: $A \equiv (L, l|l)$; $B \equiv (L, l|h)$; $C \equiv (L, h|l)$; $D \equiv (L, h|h)$; $E \equiv (H, h|l)$; $F \equiv (H, h|h)$; $G \equiv (H, l|l)$; $R \equiv (H, l|h)$. These labels correspond to the information state a player moves from when expressing a demand, including the player's privately known type, and can be used to identify players at that stage of the game. Call α_x the probability that A demands x, and similarly for the other labels: β_x for B, κ_x for C, δ_x for D, η_x for E, φ_x for F, γ_x for G, and ρ_x for R. Because labels depend on the messages exchanged, only some matches are possible: A can be matched either with another A or with a G (and similarly G can only be matched with A or with another G); D can be matched either with another D or with an F (and similarly F can only be matched with D or with another F); B can be matched with either E or C, C can be matched with either B or R, R can be matched with either C or E, and finally E can be matched with either R or B.

Characterizing demand strategies, as function of type and messages, amounts to comparing expected shares for different demands, taking into account the possible matches and the opponent's expected demand. Which demand results in a higher expected share depends on the demand strategy used by the opponent and on the posterior probabilities of the different types, given the messages. Two preliminary observations are useful: (1) Any player can guarantee herself $1-\theta$ by demanding it. (2) Given $\theta/2 > 1-\theta$ and the restriction on players never playing w, demanding $1-\theta$ is dominated for any H player (since war against an L yields θ, and war against an H yields $\theta/2 > 1-\theta$). Demands of $1-\theta$ by H players are ignored in what follows.

Thus, for example, A and G's expected shares for different demands are given by:

$$S_A(1-\theta) = 1-\theta,$$

$$S_A(1/2) = \pi_l(1-\gamma_\theta)/2 + (1-\pi_l)[(1-\alpha_\theta)/2 + \alpha_\theta(\theta/2)],$$

$$S_A(\theta) = \pi_l(1-\gamma_{1/2}-\gamma_\theta)\theta + (1-\pi_l)[\theta(1-\alpha_{1/2}-\alpha_\theta) + (\alpha_{1/2}+\alpha_\theta)(\theta/2)],$$

$$S_G(1/2) = \pi_l[(1-\gamma_\theta)/2 + \gamma_\theta(\theta/2)] + (1-\pi_l)[(1-\alpha_\theta)/2 + \alpha_\theta\theta],$$

$$S_G(\theta) = \pi_l[(1-\gamma_\theta)\theta + \gamma_\theta(\theta/2)] + (1-\pi_l)\theta,$$

where, in the absence of silence, π_l, the posterior probability that the opponent is H after the opponent has sent message l, is given by:

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

Note that if $\tau_H = 1 - \tau_L$ the messages are fully uninformative, and $\pi_l = \pi_h = q$.

In characterizing equilibria that are relevant for the lab, allowing for a small but positive probability of any message is a simple and realistic means of guaranteeing that posterior probabilities are always well-defined. In other words, we select equilibria such that any information state allowed by the structure of the game is reached with positive probability along the equilibrium path. When silence is ruled out, we impose $\tau_L \in (0, 1)$, with open bounds.

The equations corresponding to the other labels can be written in similar fashion and are not reported here.

Demand stage.

At the demand stage, given messages, the following demands are mutual best responses.

A and G: (1) G demands θ; A demands θ if $\pi_l \le (3\theta - 2)/\theta$, mixes between θ and $1-\theta$ if $\pi_l \in ((3\theta-2)/\theta, (2\theta-1)/\theta)$, and demands $1-\theta$ if $\pi_l \ge (2\theta-1)/\theta$. (2) G demands θ; A demands $1/2$ if $\pi_l \le 2\theta - 1$. (3) G demands $1/2$; A demands $1/2$ if $\pi_l \ge (2\theta-1)/\theta$.

D and F: (1) F demands θ; D demands θ if $\pi_h \le (3\theta - 2)/\theta$, mixes between θ and $1-\theta$ if $\pi_h \in ((3\theta-2)/\theta, (2\theta-1)/\theta)$, and demands $1-\theta$ if $\pi_h \ge (2\theta-1)/\theta$. (2) F demands θ; D demands $1/2$ if $\pi_h \le 2\theta - 1$. (3) F demands $1/2$; D demands $1/2$ if $\pi_h \ge (2\theta-1)/\theta$.

B, C, R and E: (1) Both E and R demand θ, B demands $1-\theta$ and C demands θ if $\pi_l \le (2\theta-1)/\theta$ and $\pi_h \ge (3\theta-2)/\theta$. (2) Both E and R demand θ, C demands $1-\theta$ and B demands θ if $\pi_h \le (2\theta-1)/\theta$ and $\pi_l \ge (3\theta-2)/\theta$. (3) Both E and R demand θ, and both B and C mix between $1-\theta$ and θ if $\pi_h \in [(3\theta-2)/\theta, (2\theta-1)/\theta]$ and $\pi_l \in [(3\theta-2)/\theta, (2\theta-1)/\theta]$. (4) Both E and R demand θ, B and C demand $1/2$ if $\pi_l \le 2\theta - 1$ and $\pi_h \le 2\theta - 1$.

Message stage

Consider now the problem for an L and an H type, choosing which message to send at the message stage. The objective is to maximize the expected share of the pie, which we now denote as $S_T(m)$ for a player of type T who sends message m. We use the symbol $\widehat{S}_Y$ to indicate the expected share of player with label Y at the allocation stage under mutual best response demand strategies. Thus:

$$\begin{aligned} S_L(l) &= [(1-q)\tau_L + q(1-\tau_H)]\widehat{S}_A + [q\tau_H + (1-q)(1-\tau_L)]\widehat{S}_B \\ S_L(h) &= [(1-q)\tau_L + q(1-\tau_H)]\widehat{S}_C + [q\tau_H + (1-q)(1-\tau_L)]\widehat{S}_D \\ S_H(h) &= [(1-q)\tau_L + q(1-\tau_H)]\widehat{S}_E + [q\tau_H + (1-q)(1-\tau_L)]\widehat{S}_F \\ S_H(l) &= [(1-q)\tau_L + q(1-\tau_H)]\widehat{S}_G + [q\tau_H + (1-q)(1-\tau_L)]\widehat{S}_R. \end{aligned} \tag{18}$$

The terms in square brackets are the probabilities of being matched with an opponent who sends message l (the first term) or h (the second term).

Equilibria

Consider the following candidate equilibria: $\{\tau_L \in (0,1),\ \tau_H \in (0,1),\ \gamma_\theta = \eta_\theta = \rho_\theta = \varphi_\theta = 1,$ $\alpha_\theta = 1-\alpha_{1-\theta} = \beta_\theta = 1-\beta_{1-\theta} =$

$= 2\left(1 - \frac{1-\theta}{\theta(1-\pi_l)}\right) \in (0,1),\ \delta_\theta = 1-\delta_{1-\theta} = \kappa_\theta = 1-\kappa_{1-\theta} = 2\left(1-\frac{1-\theta}{\theta(1-\pi_h)}\right) \in (0,1)\}$. That is, a set of equilibria indexed by τ_L and τ_H where: all H types always demand θ at the demand stage, regardless of messages; all L types mix between demanding $1-\theta$ and demanding θ at the demand stage, with strictly positive mixing probabilities that depend on the message sent; all types, L and H, send an untruthful message with positive probability. If such an equilibrium exists, then $\widehat{S}_A = \widehat{S}_B = \widehat{S}_C = \widehat{S}_D = 1-\theta$, $\widehat{S}_G = \widehat{S}_E = \pi_l(\theta/2) + (1-\pi_l)\theta$, and $\widehat{S}_F = \widehat{S}_R = \pi_h(\theta/2) + (1-\pi_h)\theta$. It follows from (18) above that randomizing between a truthful and untruthful message is indeed a best response. From the analysis of the demand strategies above, we know that the conjectured solution imposes constraints on the posterior probabilities π_h and π_l. More precisely, we require:

$$\begin{aligned} \pi_h &\in [(3\theta-2)/\theta, (2\theta-1)/\theta] \\ \pi_l &\in [(3\theta-2)/\theta, (2\theta-1)/\theta]. \end{aligned} \tag{19}$$

For any $\tau_L \in (0,1)$, ruling out silence, conditions (19) correspond to the restrictions on τ_H identified in Proposition A1 (inequalities (16), with $\sigma_H = \sigma_L = 0$).

Finally, call p the probability that an L player demands $(1-\theta)$, unconditional on message, or:

$$p \equiv 1 - \tau_L\alpha_\theta - (1-\tau_L)\delta_H.$$

Given $\alpha_\theta = 2\left(1 - \frac{1-\theta}{\theta(1-\pi_l)}\right)$ and $\delta_\theta = \left(1 - \frac{1-\theta}{\theta(1-\pi_h)}\right)$, we find:

$$p = \frac{2 - \theta(3-q)}{\theta(1-q)} = p(\theta, q).$$

The probability that an L player demands $(1-\theta)$ depends on q and θ, but not on the message sent: even when the message is informative, that is, away from the babbling line $\tau_H = 1 - \tau_L$, the mixing probabilities at the demand stage effectively nullify the information provided by the message. The probability of the opponent demanding $(1-\theta)$ does not vary with the message. Hence neither does the ex ante probability of peace, denoted by P:

$$\begin{aligned} P &= 2q(1-q)p + (1-q)^2[1-(1-p)^2] \\ &= \frac{[\theta(5-q)-2][2-\theta(3-q)]}{\theta^2}. \end{aligned}$$

The semi-pooling equilibria where types partially distinguish themselves through their messages do not have higher peace than the corresponding equilibria with babbling,[49] or in the absence of communication.

Step 2. Adding silence: $\sigma_L > 0$, $\sigma_H > 0$.

Adding silent messages does not affect the logic of the derivation above. It complicates the analysis because new information states must be considered at the demand stage, reflecting players who either received or sent (or both sent and received) a silent message. Consider for example a player labeled $A_{s2}(L, l|s)$, an L player who sent an l message and received a silent message. A_{s2} can be matched either with $A_{s1}(L, s|l)$ or with $G_{s1}(H, s|l)$ (with index $s1$ denoting a player who sent a silent message, and $s2$ denoting a player who received it). Replicating the steps above, it is not difficult to verify that mixing between $d = 1-\theta$ and $d = \theta$ is a best response for A_{s2} if G_{s1} demands θ with certainty and A_{s1} randomizes between θ (with probability α_{s1}) and $1-\theta$ (with probability $1-\alpha_{s1}$) as long as:

$$\alpha_{s1} = 2\left(1 - \frac{1-\theta}{\theta(1-\pi_s)}\right) \in [0,1]$$

[49] We say “corresponding" equilibria with babbling because we have not ruled out other equilibria where the messages are fully uninformative but are used as coordinating devices.

where π_s is the posterior probability that an opponent who sent a silent message is H. Or:

$$\pi_s = \frac{q\sigma_H}{q\sigma_H + (1-q)\sigma_L}.$$

The constraint $\alpha_{s1} \in [0,1]$ corresponds to $\pi_s \in [(3\theta - 2)/\theta, (2\theta - 1)/\theta]$, or:

$$\sigma_H \in \left[\frac{3\theta - 2}{2(1-\theta)}\left(\frac{1-q}{q}\right)\sigma_L, \frac{2\theta - 1}{1-\theta}\left(\frac{1-q}{q}\right)\sigma_L\right] \tag{20}$$

for $\sigma_L \in [0, 1-\tau_L]$.

Condition (20) must be satisfied, together with conditions (19). With $\sigma_L > 0$, $\sigma_H > 0$, and imposing $(\tau_H + \sigma_H) \in (0,1)$, the conditions amount to the boundaries on τ_H and σ_H reported in Proposition A1. The boundaries continue to include the possibility of babbling: ($\tau_H = 1 - \tau_L - \sigma_L$, $\sigma_L = \sigma_H$). Note that the open boundaries $\sigma_T > 0$, $(\tau_T + \sigma_T) \in (0,1)$ guarantee that the all posterior probabilities are well-defined.

Now consider the message choices for an L player. Taking silence into account, expected shares become:

$$\begin{aligned}
S_L(l) &= [(1-q)\tau_L + q(1-\sigma_H-\tau_H)]\widehat{S}_A + [(1-q)\sigma_L + q\sigma_H]\widehat{S}_{A_{s2}} + [q\tau_H + (1-q)(1-\sigma_L-\tau_L)]\widehat{S}_B \\
S_L(h) &= [(1-q)\tau_L + q(1-\sigma_H-\tau_H)]\widehat{S}_C + [(1-q)\sigma_L + q\sigma_H]\widehat{S}_{D_{s2}} + [q\tau_H + (1-q)(1-\sigma_L-\tau_L)]\widehat{S}_D \\
& \qquad (21) \\
S_L(s) &= [(1-q)\tau_L + q(1-\sigma_H-\tau_H)]\widehat{S}_{A_{s1}} + [(1-q)\sigma_L + q\sigma_H]\widehat{S}_{L_{ss}} + [q\tau_H + (1-q)(1-\sigma_L-\tau_L)]\widehat{S}_{Ds1}
\end{aligned}$$

where we use label D_{s1} for player $(L, s|h)$, D_{s2} for $(L, h|s)$, and L_{ss} for $(L, s|s)$. In the candidate equilibrium, all expected shares at the demand stage, conditional on messages and on best response demand strategies, equal $(1-\theta)$. Thus the player is indifferent over all three messages, and messages can be randomized. The same observation applies to an H player, who thus again is indifferent. The randomization over the messages is supported.

As above, call p the probability that an L player demands $(1-\theta)$, unconditional on message, or:

$$p \equiv 1 - \tau_L\alpha_\theta - \sigma_L\alpha_{\theta,s1} - (1 - \tau_L - \sigma_L)\delta_H$$

where $\alpha_{\theta,s1} = 2\left(1 - \frac{1-\theta}{\theta(1-\pi_s)}\right)$ is the probability with which an L player demands θ after a silent message. Given $\alpha_\theta = 2\left(1 - \frac{1-\theta}{\theta(1-\pi_l)}\right)$ and $\delta_\theta = \left(1 - \frac{1-\theta}{\theta(1-\pi_h)}\right)$, once again we find:

$$p = \frac{2-\theta(3-q)}{\theta(1-q)} = p(\theta, q).$$

As before, p does not depend on the message sent, and hence is not affected by the possibility of a silent message. As before, even informative communication has no impact on the ex ante probability of peace P:

$$P = \frac{[\theta(5-q)-2][2-\theta(3-q)]}{\theta^2}.$$

(2). Result (2) follows from the identical logic. It is not difficult to verify that, at the demand stage, all H players demanding θ and all L players demanding $1/2$ are mutual best responses if $\pi_l \leq 2\theta - 1$, $\pi_h \leq 2\theta - 1$, and, when incorporating the possibility of silence, $\pi_s \leq 2\theta - 1$. The inequalities correspond to constraints (17) in the proposition. As long as these constraints are satisfied, messages are irrelevant and mixing over messages is indeed a best response at the message stage. □

With $\theta = 0.7$, conditions (16) become:

$$\tau_H \in [\max[(1/6)(1-\sigma_L-\tau_L), 1-\sigma_H-(4/3)\tau_L], \min[(4/3)(1-\sigma_L-\tau_L), 1-\sigma_H-(1/6)\tau_L]$$
$$\sigma_H \in [(1/6)\sigma_L, (4/3)\sigma_L]$$

if $q = 1/2$, and:

$$\tau_H \in [\max[(1/3)(1-\sigma_L-\tau_L), 1-\sigma_H-(8/3)\tau_L], \min[(8/3)(1-\sigma_L-\tau_L), 1-\sigma_H-(1/3)\tau_L]$$
$$\sigma_H \in [(1/3)\sigma_L, (8/3)\sigma_L]$$

if $q = 1/3$.

The constraints corresponding to the second equilibrium, if $q = 1/3$, are reported in the text.

10 Supplemental Material

10.1 Characterization of HM Equilibria in Lemma A1

Following the first part of the proof of Lemma A1 in Appendix 8.1, 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 $q_s \in (0,1)$ 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 $p50_{ll}$.

By construction of the candidate equilibrium strategies, HM is optimizing given others' behavior and the acceptance strategies are optimal. It remains to check that sincerity is optimal for both 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:

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

If $Pr(1/2,1/2|hl)=1$, the condition becomes: $Pr(1/2,1/2|hh)Pr(H_j \text{ accepts } 1/2)\geq(\frac{1-q}{q})(\frac{2\theta-1}{1-\theta})$. But: $(\frac{1-q}{q})(\frac{2\theta-1}{1-\theta})>1 \iff q<\frac{2\theta-1}{\theta}$, 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, HM 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 L's 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".

10.2 Human Mediation with No Incentive to Walk Out

For the human mediation game, why do we set the mediators payoffs as we do, as opposed to giving the mediator no incentive to walk out (i.e. giving her a payoff of 1 if an offer is accepted and a payoff of 0 otherwise)? Our main justification is the following result. Concentrating on equilibria for which $\tau_H = 1$, $r(l,l) = (50,50)$ w.p.1, dominant acceptance strategies are taken (both players accept 70, L accepts 50, and H rejects 30), and $r(h,l) = (70,30)$ w.p. > 0, we find that there can be no non-pooling equilibrium for $q = \frac{1}{3}$.

Proposition HM*. *For $q = \frac{1}{3}$, there is no non-pooling equilibrium in which $\tau_H = 1$, $r(l,l) = (50,50)$ w.p. 1, and dominant acceptance strategies are taken (both players accept 70, L accepts 50, and H rejects 30), and $r(h,l) = (70,30)$ w.p. $>$ 0.*

To prove this proposition, we make use of the following Lemma A4 below, which is a version of Lemma A3 for this alternate specification of the mediator's payoffs. Most of the proof is identical to that of Lemma A3.

Lemma A4. *In any PBE such that $\tau_H = 1$, $r(l,l) = (50,50)$ w.p. 1, and dominant acceptance strategies are taken: (i) if $r(h,l) = (70,30)$ w.p. $>$ 0, then $\beta_l = 1$ and thus $r(h,l) = (70,30)$ is accepted w.p. 1; (ii) if $r(h,h) = (70,30)$ w.p. $>$ 0 (randomizing which player receives 30), then $\beta_h = 1$; (iii) if $r(h,l) = (70,30)$ w.p. $>$ 0, then $r(h,l) = (70,30)$ w.p. 1; (iv) if $r(h,l) = (70,30)$ w.p. $>$ 0, then $\alpha_l = 0$; and (v):*

- *HM is willing to choose $r(h,h) = w$ over $r(h,h) = (70,30)$ if and only if*

$$\tau_L = 1 \text{ or } \beta_h = 0.$$

- *HM is willing to choose $r(h,h) = w$ over $r(h,h) = (50,50)$ if and only if*

$$\tau_L = 1 \text{ and } \alpha_h = 0.$$

- *HM prefers $r(h,h) = (50,50)$ to $r(h,h) = (70,30)$ if and only if*

$$q_h^2\alpha_h^2 + 2q_h(1-q_h)\alpha_h + (1-q_h)^2 \geq (1-q_h)\beta_h,$$

  where $q_h = \frac{q}{q+(1-\tau_L)(1-q)} \in [q,1]$.

- *If $r(h,l) = (70,30)$ w.p. > 0 and $r(h,h) = (50,50)$ w.p. > 0, then H prefers to accept 50 after messaging h if and only if*

$$\tau_L \geq 1 - \frac{3}{4}(\frac{q}{1-q})\alpha_h.$$

Proof. The proofs for parts *(i)-(iv)* are exactly the same as those for parts *(i)-(iv)* of Lemma A3.

(v) For simplicity, normalize payoffs such that an accepted offer gives HM a payoff of 1 and rejection or walking out gives HM a payoff of 0. Let uW_{hh}, $u50_{hh}$, and $u70_{hh}$ be HM's expected payoffs from walking out, offering $(50,50)$, and offering $(70,30)$, respectively, after observing (h,h). These are given as

$$uW_{hh} = 0,$$

$$u50_{hh} = [q_h^2\alpha_h^2 + 2q_h(1-q_h)\alpha_h + (1-q_h)^2],$$

$$u70_{hh} = [(1-q_h)\beta_h].$$

HM is willing to choose $r(h,h) = w$ over $r(h,h) = (70,30)$ if and only if

$$\begin{aligned} 0 \geq &(1-q_h)\beta_h \iff \\ &q_h = 1 \text{ or } \beta_h = 0 \iff \\ &\tau_L = 1 \text{ or } \beta_h = 0, \end{aligned}$$

where we have used that, given $\tau_H = 1$, $q_h = 1 \iff \tau_L = 1$.

HM is willing to choose $r(h,h) = w$ over $r(h,h) = (50,50)$ if and only if

$$\begin{aligned} 0 \geq & [q_h^2\alpha_h^2 + 2q_h(1-q_h)\alpha_h + (1-q_h)^2] \iff \\ & q_h = 1 \text{ and } \alpha_h = 0 \iff \\ & \tau_L = 1 \text{ and } \alpha_h = 0, \end{aligned}$$

where, again, we have used that, given $\tau_H = 1$, $q_h = 1 \iff \tau_L = 1$. The remaining indifference condition for HM and the indifference condition for Hh to accept 50 are derived exactly as in Lemma A3. □

Proof of Proposition HM*. We search for equilibria in undominated strategies in which $\tau_H = 1$, $r(l,l) = (50,50)$ w.p. 1, $r(h,l) = (70,30)$ w.p. > 0, and $\tau_L > 0$, i.e., non-pooling equilibria. In such an equilibrium, by Lemma A4, it must be that $r(h,l) = (70,30)$ w.p. 1, $\beta_l = 1$, and $\alpha_l = 0$.

For any equilibrium in this class, HM's strategy is determined up to what she does following (h,h). The proof proceeds by going through all HM's possible strategies following (h,h) and determining if the strategy can be supported in an equilibrium with the necessary features.

Case 1: $r(h,h) = w$ w.p. 1.

From Lemma A4, in order for HM to be willing to do this, it must be that $\tau_L = 1$; and it also must be that $\beta_l = 1$. Thus, L's payoff from messaging l is such that she receives 30 against an H type and 50 against an L type for a payoff of $30q + 50(1-q) = 50 - 20q$. If instead, she lies and messages h, she will receive 0 against an H type and 70 against an L type for a payoff of $0q + 70(1-q) = 70 - 70q$. Hence, when $q = \frac{1}{3}$, the L-type prefers to lie and there is no equilibrium.

Case 2: $r(h,h) = (50,50)$ w.p. 1.

From Lemma A4, in order for HM to be willing to do this, it must be that

$$q_h^2\alpha_h^2 + 2q_h(1-q_h)\alpha_h + (1-q_h)^2 \geq (1-q_h)\beta_h.$$

If $\alpha_h = 1$, this is satisfied (the LHS equals 1), but then L's optimal strategy would involve messaging h w.p. 1, i.e. $\tau_L = 0$.

If $\alpha_h = 0$, then the condition becomes

$$(1-q_h)^2 \geq (1-q_h)\beta_h,$$

which may be satisfied for sufficiently low β_h. Can it be that $\tau_L = 1$? Since H must reject 50, this cannot be for $q = \frac{1}{3}$ since this then effectively becomes Case 1 in which h, h leads to rejection w.p. 1: the same expressions govern L's payoffs for each message, and she would prefer to lie and message h. What about $\tau_L \in (0, 1)$? This requires that $EU_L(l) = EU_L(h)$ (X refers to payoffs conditional on reaching off-path nodes, which do not matter):

$$\begin{aligned} EU_L(l) &= q \cdot \{\tau_H \cdot 30 + (1 - \tau_H)X\} + (1 - q)\{\tau_L \cdot 50 + (1 - \tau_L) \cdot 30\} \\ &= 30 + 20\tau_L - 20\tau_L q \end{aligned}$$

and

$$\begin{aligned} EU_L(h) &= q \cdot \{\tau_H \cdot 50 \cdot \alpha_h + (1 - \tau_H)X\} + (1 - q)\{\tau_L \cdot 70 + (1 - \tau_L) \cdot 50\} \\ &= q50 \cdot \alpha_h + 50 + 20\tau_L - 50q - 20\tau_L q, \end{aligned}$$

and thus

$$\begin{aligned} EU_L(l) = EU_L(h) &\iff \\ 30 + 20\tau_L - 20\tau_L q = q50 \cdot \alpha_h + 50 + 20\tau_L - 50q - 20\tau_L q &\iff \\ 50q = q50 \cdot \alpha_h + 20 &\iff \\ \alpha_h = \frac{50q - 20}{50q} = \begin{cases} \frac{1}{5} & q = \frac{1}{2} \\ -\frac{1}{5} & q = \frac{1}{3}, \end{cases} \end{aligned}$$

and this is not possible for $q = \frac{1}{3}$.

If $\alpha_h \in (0, 1)$, then it must be that H who messages h is indifferent between accepting or rejecting 50. Since $r(h, h) = (50, 50)$ is offered w.p. > 0, the condition for this is $\tau_L = 1 - \frac{3}{4}(\frac{q}{1-q})\alpha_h$, which implies that $\tau_L \in (0, 1)$, and so it must be that $EU_L(l) = EU_L(h)$. But the above expression shows that this is not possible for $q = \frac{1}{3}$

Case 3: $r(h, h) = (70, 30)$ w.p. 1.

That $r(h, h) = (70, 30)$ w.p. > 0 implies that $\beta_h = 1$ from Lemma A4. The lemma then implies that, in order for HM to be willing to offer $r(h, h) = (70, 30)$, it must be that

$$(1 - q_h) \geq q_h^2 \alpha_h^2 + 2q_h(1 - q_h)\alpha_h + (1 - q_h)^2.$$

This is satisfied if α_h is sufficiently low, which we can set since H who messages h is never offered 50, and so low α_h is optimal given some beliefs (i.e. if the opponent is L with sufficiently high probability). Is it possible that L is willing to sometimes message l so that $\tau_L > 0$? This requires that $EU_L(l) \geq EU_L(h)$ (X refers to payoffs conditional on reaching off-path nodes, which do not matter):

$$\begin{aligned} EU_L(l) &= q \cdot \{\tau_H \cdot 30 + (1 - \tau_H)X\} + (1 - q)\{\tau_L \cdot 50 + (1 - \tau_L) \cdot 30\} \\ &= 30 + 20 \cdot \tau_L - 20 \cdot \tau_L q \end{aligned}$$

and

$$\begin{aligned} EU_L(h) &= q \cdot \{\tau_H(\frac{1}{2}30 + \frac{1}{2}0) + (1 - \tau_H)X\} + (1 - q)\{\tau_L \cdot 70 + (1 - \tau_L) \cdot (\frac{1}{2}30 + \frac{1}{2}70)\} \\ &= -35q + 50 + 20\tau_L - 20\tau_L q, \end{aligned}$$

and thus

$$\begin{aligned} EU_L(l) \geq EU_L(h) \iff \\ q \geq \frac{4}{7}, \end{aligned}$$

and this is false. Hence, any such equilibrium involves $\tau_L = 0$.

Case 4: HM mixes between $r(h, h) = w$, $r(h, h) = (50, 50)$, and $r(h, h) = (70, 30)$.

From Lemma A4, in order for HM to be willing to do this, it must be that

$$\tau_L = 1 \text{ and } \alpha_h = 0.$$

Can it be that $\tau_L = 1$? Since H must always reject 30 and 50, this cannot be for $q = \frac{1}{3}$ since this then effectively becomes Case 1 in which (h, h) leads to rejection w.p. 1: the same expressions govern L's payoffs for each message, and she would prefer to lie and message h.

Case 5: HM mixes between $r(h,h) = w$ and $r(h,h) = (50,50)$.

From Lemma A4, in order for HM to be willing to do this, it must be that

$$\tau_L = 1 \text{ and } \alpha_h = 0,$$

Can it be that $\tau_L = 1$? Since H must reject 50, this cannot be for $q = \frac{1}{3}$ since this then effectively becomes Case 1 in which (h,h) leads to rejection w.p. 1: the same expressions govern L's payoffs for each message, and she would prefer to lie and message h.

Case 6: HM mixes between $r(h,h) = w$ and $r(h,h) = (70,30)$.

From Lemma A4, in order for HM to be willing to do this, it must be that

$$\tau_L = 1 \text{ or } \beta_h = 0$$

Can it be that $\tau_L = 1$? Since H must always reject 30, this cannot be for $q = \frac{1}{3}$ since this then effectively becomes Case 1 in which (h,h) leads to rejection w.p. 1: the same expressions govern L's payoffs for each message, and she would prefer to lie and message h.

Case 7: HM mixes between $r(h,h) = (50,50)$ and $r(h,h) = (70,30)$.

That $r(h,h) = (70,30)$ w.p. > 0 implies that $\beta_h = 1$ from Lemma A4. The lemma then implies that, in order for HM to be willing to mix between $r(h,h) = (50,50)$ and $r(h,h) = (70,30)$, it must be that

$$(1-q_h) = q_h^2\alpha_h^2 + 2q_h(1-q_h)\alpha_h + (1-q_h)^2.$$

If $\alpha_h = 0$, then this is satisfied only if $q_h = 1 \iff \tau_L = 1$. But then this cannot be part of an equilibrium for $q = \frac{1}{3}$ since this then effectively becomes Case 1 in which (h,h) leads to rejection w.p. 1: the same expressions govern L's payoffs for each message, and she would prefer to lie and message h.

If $\alpha_h = 1$, then the RHS of the above equals 1, and so it cannot be satisfied given $\tau_H = 1 \implies q_h > 0$.

Hence, we require that $\alpha_h \in (0,1)$. Since $r(h,h) = (50,50)$ is offered w.p. > 0, the condition that must be satisfied is

$$\tau_L = 1 - \frac{3}{4}(\frac{q}{1-q})\alpha_h,$$

which implies

$$q_h = \frac{q}{q + (1 - \tau_L)(1 - q)} = \frac{1}{1 + \frac{3}{4}\alpha_h} = \frac{4}{4 + 3\alpha_h}.$$

Plugging into HM's indifference condition, letting $q = \frac{1}{3}$, and numerically solving yields

$$\alpha_h^* \approx 0.300 \text{ and } \tau_L^* \approx 0.887.$$

Since L must be indifferent between both messages, a necessary condition for equilibrium is that $p50_{hh} \in (0, 1)$, the probability with which $r(h, h) = (50, 50)$, is such that $EU_L(l) = EU_L(h)$ (X refers to payoffs conditional on reaching off-path nodes, which do not matter):

$$\begin{aligned} EU_L(l) &= q \cdot \{\tau_H \cdot 30 + (1 - \tau_H)X\} + (1 - q)\{\tau_L^* \cdot 50 + (1 - \tau_L^*) \cdot 30\} \\ &= 30 + 20 \cdot \tau_L^* - 20 \cdot \tau_L^* q \end{aligned}$$

and

$$\begin{aligned} EU_L(h) =& q \cdot \{\tau_H(p50_{hh} \cdot 50 \cdot \alpha_h^* + (1 - p50_{hh})(\frac{1}{2}30 + \frac{1}{2}0)) + (1 - \tau_H)X\} \\ &+ (1 - q)\{\tau_L^* \cdot 70 + (1 - \tau_L^*)(p50_{hh} \cdot 50 + (1 - p50_{hh})(\frac{1}{2}30 + \frac{1}{2}70))\} \\ =& p50_{hh}\{50\alpha_h^* - 15\}q + \{50 - 35q + 20\tau_L^* - 20\tau_L^* q\}, \end{aligned}$$

and thus

$$EU_L(h) - EU_L(l) = p50_{hh}\{50\alpha_h^* - 15\}q + \{20 - 35q\},$$

which is greater than zero for all $p50_{hh} \in [0, 1]$. This means that there does not exist $p50_{hh} \in (0, 1)$ such that $EU_L(l) - EU_L(h)$, and so there is no such equilibrium. □

10.3 Additional Experimental Results

10.3.1 Sincerity and Information Transmission: Kullback-Leibler Measures

As noted in the text, how much information a message transmits depends on the use of that same message by the opposite type. We can use the Kullback-Leibler (KL) measure of dispersion to generate a summary indicator of the impact of a message on the posterior probability of a given type, relative to the prior, taking into account the use of the message by both types. For the two messages h and

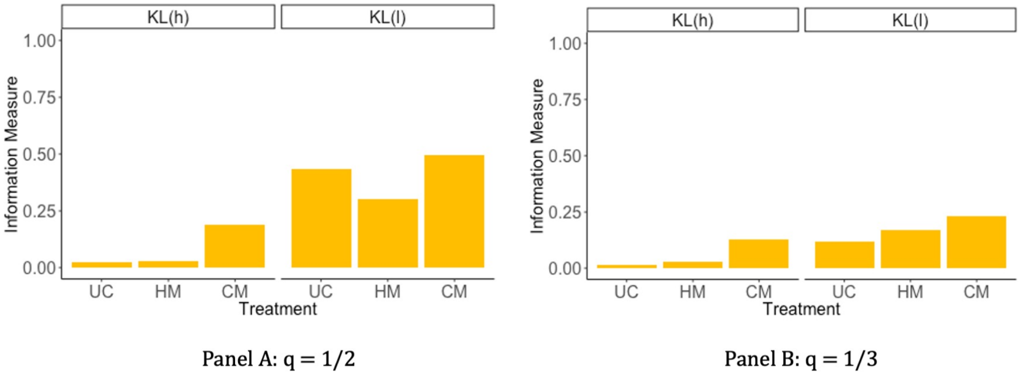
l, the respective KL measures are:

$$KL(h) = \Pr(H|h) \log \left(\frac{\Pr(H|h)}{\Pr(H)} \right) + \Pr(L|h) \log \left(\frac{\Pr(L|h)}{\Pr(L)} \right)$$

$$KL(l) = \Pr(H|l) \log \left(\frac{\Pr(H|l)}{\Pr(H)} \right) + \Pr(L|l) \log \left(\frac{\Pr(L|l)}{\Pr(L)} \right).$$

KL measures are always non-negative and equal 0 when the posterior equals the prior (no information has been conveyed). In our setting, maximal values are $-\log(q)$ for $KL(h)$ and $-\log(1-q)$ for $KL(l)$.

Figure 14 reports, for the three treatments and the two parameterizations, the corresponding KL measures for messages l and h, expressed as fractions of the maximum value for each parameterization and averaged over the relevant sessions.



Panel A: q = 1/2

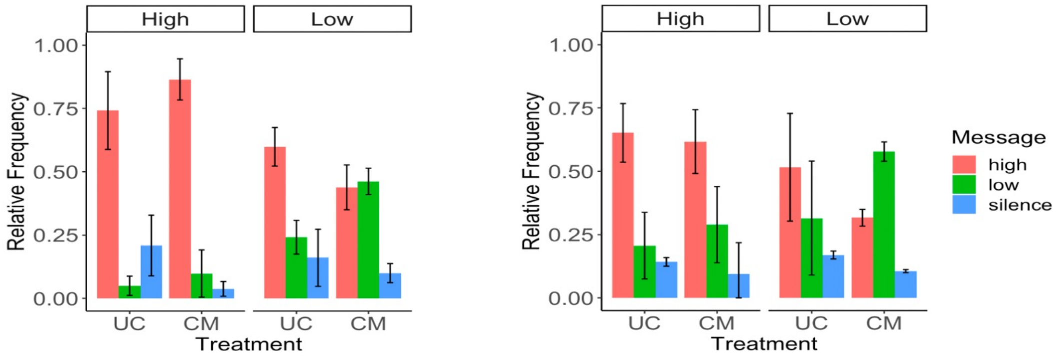
Panel B: q = 1/3

Figure 14: KL measures.

It remains true that the treatment conveying most information is CM, although the lesson from the KL measures is more nuanced than Figure 2 in the text and Figure 15 below suggest. The high sincerity of the H types does not translate into high information from the h message, since the same message is also used by the L types. Message l on the other hand, is more informative even though sincerity is less common among L's because few are the H types who send message l. The importance of the interaction in the use of the messages between the two types becomes very clear when comparing the two parameterizations. Even though L's tend to be more sincere with $q = 1/3$, the more common use of message l by H types severely reduces the information transmitted by the messages, relative to the $q = 1/2$ parameterization.

10.3.2 Sincerity and Peace in UC and CM, between Subjects

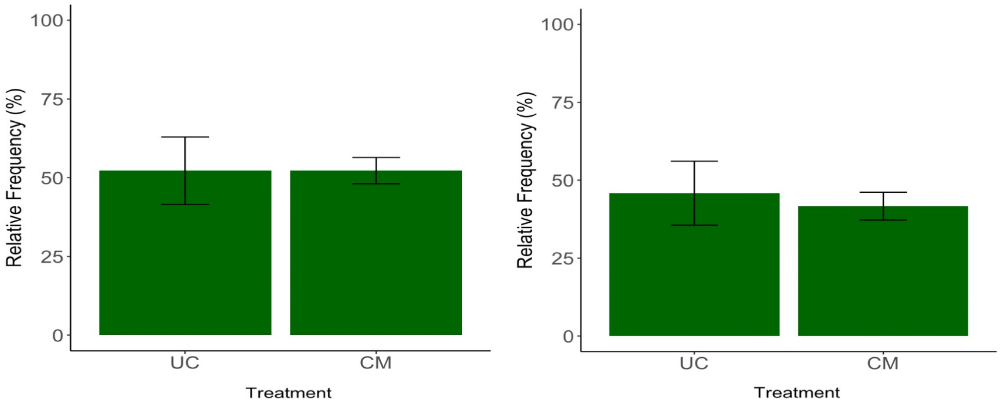
We report in Figures 15 and 16 evidence on sincerity and the frequency of peace in UC and CM, comparing sessions where the two treatments are played by subjects in rounds 11-30, i.e. just after the NC treatment, and with the same experience.



Panel A: $\mathrm{q} = 1/2$

Panel B: $\mathrm{q} = 1/3$

Figure 15: UC and CM: sincerity with equal experience.



Panel A: $\mathrm{q} = 1/2$

Panel B: $\mathrm{q} = 1/3$

Figure 16: UC and CM: frequency of peace with equal experience.

10.3.3 Peace Regressions with Full Interactions Terms and by Treatments

We report in Table 3 the results of a linear regression of the frequency of peace on treatment, parameterization, types-pair, order, and round as in Table 2 in the text but with the full set of interaction terms. The substantive results are unchanged but the regression clarifies two points: first, HM's lower frequency of peace is particularly noticeable with $L - L$ pairs, where the negative effect is not only very significant statistically but also large quantitatively; second, it is $L - L$ pairs again, and only $L - L$ pairs, that drive the small but positive effect of learning identified in Table 2.

In the text, we concluded the description of Table 2 by remarking that the frequency of peace was significantly higher under $q = 1/2$ than under $q = 1/3$. The result is shown more transparently in regressions specialized by treatment, Table 4 below. Recall that although higher frequency of peace when the probability of H type realizations is higher is superficially counterintuitive, the result is in line with the theory. It is predicted in the HMS equilibrium under CM and in all the equilibria our analysis selects under both HM and UC.

	Dependent variable:
	Peace
HM Treatment	0.082
	(0.032)
CM Treatment	−0.028
	(0.044)
Order 2	0.066
	(0.064)
$q = 1/2$	0.154
	(0.050)
H-*L* pair	0.322
	(0.057)
L-*L* pair	0.635
	(0.072)
Round	0.0002
	(0.001)
HM Treatment × *H*-*L* pair	−0.129
	(0.025)
CM Treatment × *H*-*L* pair	0.011
	(0.038)
HM Treatment × *L*-*L* pair	−0.301
	(0.037)
CM Treatment × *L*-*L* pair	0.016
	(0.057)
Order 2 × *H*-*L* pair	−0.082
	(0.048)
Order 2 × *L*-*L* pair	−0.050
	(0.075)
$q = 1/2$ × *H*-*L* pair	0.066
	(0.046)
$q = 1/2$ × *L*-*L* pair	−0.002
	(0.059)
Round × *H*-*L* pair	0.0002
	(0.001)
Round × *L*-*L* pair	0.002
	(0.001)
Constant	0.018
	(0.064)
Observations	4,320
R^2	0.206
Adjusted R^2	0.203
Residual Std. Error	0.446 (df = 4302)

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 at the session level.

Table 3: Peace with Interactions.

	Dependent variable:		
		Peace	
	(UC)	(HM)	(CM)
Order 2	0.238	0.004	-0.110
	(0.153)	(0.081)	(0.121)
$q = 1/2$	0.078	0.099	0.083
	(0.050)	(0.081)	(0.040)
Round	-0.003	-0.003	-0.001
	(0.002)	(0.003)	(0.002)
Constant	0.516	0.556	0.553
	(0.067)	(0.137)	(0.163)
Observations	1,440	1,440	1,440
R^2	0.014	0.014	0.012
Adjusted R^2	0.012	0.012	0.010
Residual Std. Error (df = 1436)	0.496	0.495	0.498

The excluded category in the regression is $q = 1/3$ under Order 1. Standard errors are clustered at the session level.

Table 4: Peace by Treatment.

10.3.4 The No Communication (NC) rounds

All subjects started a session by playing 10 rounds of the unmediated treatment without any communication. After being matched in pairs and independently assigned types, each subject submitted a demand in the set $\{1-\theta, 1/2, \theta, w\}$; if the two demands were compatible, the resource was split accordingly, if not, conflict followed, the resource shrank to θ and was split according to the subjects' types–$\theta/2$ to each if the types were equal, θ to H and 0 to L otherwise.

These initial 10 rounds were always played before any other treatment and had the goal of familiarizing subjects with the computer interface and the bargaining game, in a simpler environment. For completeness, we briefly report here the theoretical results for the NC game as well as the main regularities observed in the data.

The UC equilibrium demand strategies described in Section 9.3 are equilibrium demand strategies under NC if the posterior probability of the opponent being H is set equal to the prior, i.e. to q. Denoting by $\delta_d(T)$ the probability that type T demands d, and by P the ex ante probability of peace, the equilibria under NC are as follows:

$\underline{q = 1/2}$. *There is a unique equilibrium:* $\delta_{70}(H) = 1; \delta_{70}(L) = 0.29$ *and* $\delta_{30}(L) = 0.71$*;* $P = 0.586$*.*

$\underline{q = 1/3}$. *There are three equilibria: (1)* $\delta_{70}(H) = 1; \delta_{50}(L) = 1;$ $P = 0.444$. (2) $\delta_{70}(H) = 1; \delta_{70}(L) = 0.71$ *and* $\delta_{30}(L) = 0.29$*;* $P = 0.345$. *(3)* $\delta_{70}(H) = 1; \delta_{70}(L) = 0.33$, $\delta_{50}(L) = 0.38$, *and* $\delta_{30}(L) = 0.29$*;* $P = 0.409$.

The equilibrium under $q = 1/2$, as well as the first two equilibria under $q = 1/3$ can be obtained from the UC equilibria discussed in Section 9.3. We have not attempted to characterize a UC equilibrium under $q = 1/3$ that would parallel the third equilibrium above.

The frequency of peace observed in the data is 0.556 under $q = 1/2$ (with s.e.'s clustered at the session level, the 95 percent CI is $[0.454, 0.657]$) and 0.528 under $q = 1/3$ (with $CI = [0.468, 0.587]$); thus peace fits the prediction quite well for $q = 1/2$ and is higher than expected under $q = 1/3$. (The predicted treatment effect, with higher peace under $q = 1/2$, is observed in the data, but with large confidence intervals). On the other hand, as shown in Table 9, demand strategies in the lab align qualitatively with the first equilibrium under $q = 1/3$ but deviate from predictions otherwise, in great part because in no other equilibrium does the L type demand 50 with high probability, a regularity we instead see consistently in the data.

		$q = 1/2$						$q = 1/3$			
demand		70	50	30	w	demand		70	50	30	w
H	data	0.66	0.29	0.003	0.04	H	data	0.80	0.14	0	0.06
	equil	1	0	0	0		equil1,2,3	1	0	0	0
L	data	0.07	0.63	0.30	0	L	data	0.10	0.85	0.04	0.01
	equil	0.29	0	0.71	0		equil1	0	100	0	0
							equil2	0.71	0	0.29	0
							equil3	0.33	0.38	0.29	0

Table 9: Data from the initial NC rounds.

10.3.5 Human Mediator Dataset

Table 10 gives the entire count data for the HM treatment. The left panels give the human mediator's recommendations following message pairs, and the right panels give the messages sent by each type.

$q = 1/2$

	hh	hl	hs	ll	ls	ss	Hh	Hl	Hs	Lh	Ll	Ls
$50-50$	209	53	33	45	6	4	565	93	45	410	277	50
$30-70$	2	0	0	1	5	0						
$70-30$	1	166	15	0	0	0						
w	177	28	22	7	6	0						

$q = 1/3$

	hh	hl	hs	ll	ls	ss	Hh	Hl	Hs	Lh	Ll	Ls
$50-50$	94	74	22	111	23	2	315	132	35	417	471	70
$30-70$	2	5	1	1	2	0						
$70-30$	0	142	9	0	0	0						
w	97	70	23	22	19	1						

Table 10: Data from the HM treatment.

10.4 Experimental Procedures and Instructions

Upon entering the lab, subjects were seated at random computer posts, divided by partitions; each subject was identified exclusively by a randomly assigned id and all communication among subjects took place exclusively via computers. Subject ids were private and not visible to other subjects. After subjects were seated and consent forms were signed, the experimenter read the instructions aloud and showed images of the experimental screenshots, answering aloud and publicly any question that did arise. We reproduce here instructions and screenshots for a representative $q = 1/2$, Order 1 session.

MEDIATION INSTRUCTIONS

Four parts: NC, UC, M, MC.

$q = 1/2$; $\theta = 0.7$.

(Payoffs for HM: M=60, W=40, m=20).

Make yourself comfortable, put your phones away, and please don't talk or use the computer. Thank you for agreeing to participate in this experiment.

You will be paid for your participation privately and in cash, at the end of the experiment. Your earnings during the experiment are denominated in POINTS. For this experiment every 100 POINTS earns you 10 DOLLARs. The experiment will consist of multiple rounds. At the end, five rounds will be selected randomly, and you will be paid the sum of your earnings over those five rounds. Pay attention to each round because it may well end up being one of those for which you will be paid.

If you have any questions during the instructions, please raise your hand.

The experiment studies a game of negotiation: you will be matched with another person, and the two of you will decide how to share a resource worth 100 points. In case of disagreement, the resource shrinks to 70 points (think of the 30 points lost as time and resources wasted to disagreement). You will be randomly assigned types, High or Low, and how the resource is divided in case of disagreement will depend on your types.

I will describe each part of the experiment before it starts.

PART 1

We begin with PART 1.

At the start of each round, the computer will assign you a type, which, as we said, can be either High or Low. The two types are equally probable: each person is likely to be H with probability 1/2, and L with probability 1/2.

You will see a screenshot like this: [SCREENSHOT ON TYPE]

Here, as at several other points during the experiment, you will move to the next screen by clicking the Continue button. Please remember to do so.

After having been assigned your type, you will be randomly matched with another person in the room. You will not know which person you are matched to, nor will you know the person's type. Knowing your type does not give you any information about your match's type. All you know is that he or she is equally likely to be H or L with probability 1/2 each. Your type and your match's type matter because they affect how the resource is shared in case of Disagreement.

After having been informed of your assigned type, you will be asked to say how much of the resource you demand for yourself. Remember that the resource is worth 100. You can ask for 30, 50, 70, or you can Walk Out of the negotiation.

- If your demand and the demand of your match are compatible (i.e. do not sum to more than 100), then they will satisfied. You will receive what you asked for, and the round will end.
- If the two demands are instead incompatible (they sum to more than 100), or if one of you Walks Out, then there is Disagreement. The resource shrinks from 100 to 70 points. The reduced resource is then allocated automatically by the computer. If one of you is H and the other is L, then H receives the full 70 points, and L receives 0. If both of you are H, or both of you are L, then each receives one half of the reduced resource, that is, 35 points. This will conclude the round.

The screen where you express your demand will look like this:

[SCREENSHOT: NO COMMUNICATION DEMAND]

Notice that you have a reminder of your type on the upper left corner.

Disagreement occurs if either of you chooses W (Walk Out), or if the two of you choose (70, 50), or (70, 70). Remember: If there is disagreement, the resource shrinks from 100 to 70 points.

After the two demands have been submitted, you will be told your match's demand; whether there is Agreement or Disagreement, and your payoff for the round.

If there is Agreement, your payoff will equal your demand. Your screen will look like this [SCREENSHOT: OUTCOME WITH AGREEMENT].

If there is disagreement, your payoff will depend on your type and your match's type. Your screen will look like this [SCREENSHOT: OUTCOME WITH DISAGREEMENT]. In this example, you asked for 70 and your match asked for 50 points. The two demands were incompatible, and the resource shrank to 70 points in total. Your payoff consists of 0 points which indicates that your match is of type H and you are of type L.

This will conclude the round. We will then move to the next round: you will again be assigned a type randomly (H or L with equal probability of 1/2), and will be matched randomly with another person in the room. The type you were in round 1 or the person you were matched with do not influence in any way the type you are assigned in round 2 or your new match. The experiment will then continue as described earlier.

The REMINDER slide summarizes this part of the experiment.

Are there any questions?

We will begin with two practice rounds. You will not be paid for these rounds, whose purpose is only to familiarize yourself with the computer interface and the rules of the experiment.

[OPEN ZTREE; copy program]

Please double-click on the icon marked Leaf16 on your desktop. If asked, click RUN.

If you have any questions from now on, raise your hand, and an experimenter will come and assist you.

RUN PRACTICE ROUNDS: [RUN; START TREATMENT]

We have now concluded the practice rounds. Are there any questions? Remember that you will not be paid for these rounds.

CLOSE THE TREE

Please click Alt F4. Then double-click on the icon marked Leaf16 and if asked click RUN.

We will now begin the experiment. The first part will last 10 rounds.

[RUN; START TREATMENT]

PART 2

We will now move to the second part of the experiment. Part 2 will run in a similar fashion to part 1. At the start of each round, the computer will again assign you a type, High or Low, with equal probability of $\frac{1}{2}$ each. You will again be matched randomly with another person in the room, whose type you will not know.

Now, unlike in Part 1, after types are assigned and matches are made, you will be asked to send a message to your match, communicating your type. You have three options: High, Low, or Silence. You can be truthful, or not truthful, as you choose, or you can be silent. The screen you will see will look like this:

[SCREENSHOT: SEND MESSAGE] As before, in the upper blue strip is a reminder of your type.

You will then receive the message sent by your match, which again can be either H or L or S. After having seen the message, you will be asked to say how much of the resource you demand for yourself. Remember that the resource is worth 100 points. As in Part 1, you can ask for 30, 50, 70, or you can Walk Out of the negotiation. Payoffs will work exactly as in the previous round: you will receive what you asked if the two demands do not sum up to more than 100 (and thus there is Agreement); if the demands sum up to more than 100, there is Disagreement, the resource shrinks to 70 points and is allocated according to your type and the type of your match.

The only difference with respect to Part 1 is your ability to send a message communicating your type before deciding on your demands.

The screen where you express your demand will look like this:

[SCREENSHOT: DEMAND] Note that blue strip at the top now reminds you both of your type and of the message you have sent. The screen also communicates to you the message your partner has sent.

After the two demands have been submitted, you will be told your match's demand; whether there is Agreement or Disagreement, and your payoff for the round.

This will conclude the round. We will then move to the next round: you will again be assigned a type randomly (H or L, each with equal probability 1/2), and will be matched randomly with another person in the room. The experiment will then continue as described earlier.

The Reminder slide will remain projected to remind you of the rules.

Part 2 will last 20 rounds.

Please move the cursor to the top left corner of your screen. Click and the Continue button will appear at the bottom right corner. Click Continue and begin Part 2.

PART 3.

We will now move to the third part of the experiment.

At the start of each round, you will be matched randomly in groups of three people. One person in the group will be called Mediator. The Mediator receives confidential messages and makes recommendations on how the other two people in the group—who will be called the two Players—are to share the resource. For convenience, the two Players will sometimes be identified as Player 1 and Player 2, but 1 and 2 are just labels with no other meaning.

The computer will tell you if you are the Mediator or a Player.

After the match has occurred, the two Players will be randomly assigned a type. As before, each type can be either H or L with equal probability, and which type is assigned to one Player has no influence on the type assigned to the other Player. If you are a Player, you will know your own type, but will not know the other Player's type. If you are the Mediator, you will not know the type of either Player. Everyone knows that a Player is assigned type H or L with equal probability of $\frac{1}{2}$ each.

After matches are made and roles and types are assigned, if you are a Player, you will be asked to send a message communicating your type, as you did in Part 2. As before, you have three options: High, Low, or Silence. The difference is that now you will send the message to the Mediator, and not to the other Player in your group. As before you can be truthful, or not truthful, or you can be Silent.

The screen will look like this:

[SCREENSHOT: SEND MESSAGE]

The message you send to the Mediator is confidential and will not be seen by the other Player.

Once the two messages are received by the Mediator, the Mediator can make a recommendation on how to share the resource, or can choose to Walk Out of the mediation.

- If the Mediator makes a recommendation and both Players accept it, then there is Agreement, the resource is shared according to the recommendation, and the Mediator earns 60 points.
- If one or both Players reject the recommendation, then there is Disagreement, the resource shrinks to 70 points and is allocated by the computer to the two Players according to their type, as in Parts 1 and 2. In case of Disagreement, the Mediator's payoff is 20 points.
- If the Mediator Walks out of the negotiation, the Disagreement scenario is triggered automatically: the resource shrinks, and the Players' payoffs depend on their type, as in the regular Disagreement case. However if Disagreement is triggered by the Mediator Walking out, the Mediator's payoff is 40 points (as opposed to 20 when Disagreement comes from the Players rejecting the Mediator's recommendations).

The reminder slide that remains projected during this part of the experiment will remind you of the rules.

Note that the Mediator can make a recommendation but has no power to force the Players to accept it.

The Mediator's screen will look like this:

[SCREENSHOT: MEDIATOR'S CHOICE]. The screen shows the two messages received from the two Players, and the options the Mediator has for a feasible recommendation. The first number indicates the amount recommended for Player 1, and the second the amount recommended for Player

2. The choices are (50,50), (30,70), (70, 30). Alternatively, the Mediator can choose to Walk Out of the mediation task.

The Mediator's choice is then transmitted to the two Players.

If the Mediator has chosen to Walk Out, then each Player will see a screen like this:

[SCREENSHOT: MEDIATOR WALKED OUT].

At the same time, the Mediator will also see a screen repeating the decision to Walk Out and reporting the Mediator's corresponding payoff. [SCREENSHOT: YOU WALKED OUT].

If the Mediator has made a recommendation, each Player's screen will look like this: [SCREENSHOT: PLAYER'S RESPONSE TO THE MEDIATOR'S PROPOSAL]. The Player is asked whether to accept or reject the recommendation.

Each Player is then told whether the other Player accepted the recommendation, and the final outcome of the mediation, including the Player's payoff for the round. [SCREENSHOT: OUTCOME FOR PLAYER, AGREEMENT].

At the same time, the Player's decisions and the outcome are communicated to the Mediator. The Mediator is also reminded of the messages received and the recommendation made. [SCREENSHOT: OUTCOME FOR MEDIATOR, AGREEMENT]. Because the outcome is Agreement, the Mediator earns 60 points.

This concludes the round. We will then move to the next round, where groups of three will again be formed randomly, and roles will be assigned randomly. Although roles are assigned randomly and groups are formed randomly, each of you will be Mediator for the same number of rounds. Types are then assigned, again randomly, with each Player being of type H or L with equal probability of $\frac{1}{2}$ each. The experiment will then continue as just described.

[SCREENSHOT: REMINDER SLIDE SUBJECT MEDIATOR]

Part 3 will last 30 rounds.

Are there any questions?

Please move the cursor to the top left corner of your screen. Click and the Continue button will appear at the bottom right corner. Click Continue and begin Part 3.

PART 4

Part 4 is almost identical to Part 3. The only difference is that the Mediator is played by the computer.

As in Part 3, the two Players in each group send their messages to the Computer-Mediator, the

Mediator chooses whether to Walk Out or to make a recommendation, and each Player decides whether to accept or to reject the Mediator's recommendation.

If either the Mediator Walks Out or one or both Players reject the Mediator's recommendation, then there is Disagreement, the resource shrinks to 70 points and is allocated according to Players' types (divided equally if the Players are of the same type, given fully to the H type if the two Players have type H and L).

If the Mediator makes a recommendation and both Players accept it, then there is Agreement, the recommendation is implemented, each Player earns the corresponding points.

The Computer Mediator follows the following plan:

If the two messages are (L, L), it recommends (50, 50).

If they are (H, L), it recommends either (70, 30) with probability 5/8 or (50, 50) with probability 3/8.

If they are (H, H), it recommends either (50, 50) with probability 1/2 or Walks Out with probability 1/2.

If the computer receives a Silent message from a player, it interprets it interprets it according to the likely frequency of each type—as an H with probability 1/2 and an L with probability 1/2. Thus if, for example, the two messages are (S,L), the computer reads them as (L,L) with probability 1/2 (and acts accordingly) or as (H,L) with probability 1/2 (and acts accordingly).

[SCREENSHOT COMPUTER MEDIATOR PLAN]

This screenshot will remain up throughout Part 4 to remind you of the Computer Mediator plan.

After each round, you will be rematched randomly with another player, and types will be reassigned.

Part 4 will last for 20 rounds.

Are there any questions?

Please move the cursor to the top left corner of your screen. Click and the Continue button will appear at the bottom right corner. Click Continue and begin Part 4.

[Before the end of the last round: remind them to remain with the final screen with their earnings].

END OF THE EXPERIMENT

This is the end of the experiment. You should now see a popup window, which displays your total earnings. Please divide the number of points by 10, round up to the nearest dollar, and record this on your payment receipt sheet. Please also enter $10.00 on the show-up fee row. Add your earnings and the show-up fee and enter the sum as the total. Finally, please record your Computer ID on the form. Add some not intelligible signature. When you are done, click “Continue”.

[Run Questionnaire]

We will pay each of you in private in the next room in the order of your computer numbers. Please do not use the computer; be patient, and remain seated until we call you to be paid. Do not converse with the other participants. Thank you for your cooperation.

[SAVE DATA and erase from folder]