On the Fragility of Mediation: Theory and Experimental Evidence*

Alessandra Casella† Evan Friedman‡ Manuel Perez Archila§

December 11, 2025

Abstract

Mediation of disputes is increasingly common, often implemented by computer-run algorithms. We test the efficacy of a theoretically optimal mediation algorithm in an experiment where two subjects, uncertain about each other's strength, negotiate how to share a resource. The subjects send cheap talk messages to one another (under direct communication) or to the computer mediator (under mediated communication), before expressing demands or receiving the mediator's non-binding recommendation. While messages to the mediator are more sincere, we find that peaceful resolution is not more frequent. The theoretical analysis shows that mediation is fragile precisely when it is most promising. When the optimal equilibrium improves over direct communication, any deviation from full truthfulness, no matter how small, causes a discontinuous downward jump in the probability of agreement.

*We thank Vasiliki Skreta, co-editor of this journal, three anonymous referees, and participants at numerous seminars and conferences 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 indebted to Shuhua Si for help running experimental sessions. We are extremely grateful to Massimo Morelli for his advice and financial support through ERC grant 694583. The experiment was approved under Columbia IRB Protocol AAAR5160.

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

‡Paris School of Economics, evan.friedman@psemail.eu

§Princeton University, mp1278@princeton.edu

1 Introduction

With applications ranging from family disputes to corporate law, from labor relations to international conflicts, mediation is increasingly advocated by psychologists, lawyers and judges, lay people, and professionals specializing in its craft. Alternative Dispute Resolution procedures have become favored solutions to the time delay and costs of judicial processes and have flourished 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 surprising. The mediator is an impartial third party who has neither independent resources, nor superior information, nor enforcement power. On what basis can the mediator's presence be helpful? Part of the answer is likely to 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's strength or weakness. The uncertainty makes them willing to accept a recommendation they would otherwise reject. The final result is a higher frequency of peaceful resolutions than what the two sides could 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 participants in the experiment, as parties in a dispute. We thus offer a test of a specific form of (theoretically optimal) algorithmic mediation. Our 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.

[1]The number of legal cases brought to trial in US courts experienced a startling 60% decline from the mid 80s 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 June 14, 2024) offers a panoramic view.

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 non-binding recommendation or refusing to mediate. Agreement is reached 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, there are well-known algorithms, similar to ours, that help to allocate resources between two parties with conflicting claims. They aim at fair and envy-free outcomes (for instance, when applied to disputes stemming from a divorce) 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, these 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. Contrary to the theory, however, mediation does not increase the frequency of agreement. In the lab, mediation does not fulfill its promise.

Having established this result, we devote most of the paper to understanding its causes. The experimental data lead us to our most novel theoretical finding: the fragility of the obfuscation mechanism. The optimal equilibrium involves full sincerity of messages, but under obfuscation, equilibria with even the smallest deviation from full sincerity imply a discontinuous jump downward in the probability of peace, a jump 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 hence the mediator's

[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.

recommendations reveal the parties' messages. The problem is that the theoretical superiority of mediation relies on obfuscation. Whenever optimal mediation involves obfuscation, the best equilibrium improves over direct communication, but is fragile. Whenever optimal mediation does not involve obfuscation, the best equilibrium is robust to some deviations from truthfulness, 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 has not 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, single 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 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; Ljungqvist 1993). Rigorous experiments are scarce, but the available results are consistent with what we find. Optimal mechanisms with randomization fall short: while they increase sincerity, they do not induce full truthfulness, and lead to outcomes that are broadly comparable to simple direct elicitation (John et al. 2018; Blume et al. 2019; Blume et al. 2023). 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.

The fragility of the obfuscation equilibrium is our most novel finding, but we also find that the optimal mechanism has 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 for actions to approximate individual best responses, nevertheless impacts 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. This literature makes clear that the comparison is sensitive to the details of the game: how

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

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 (see Fanning (2021) and the related literature 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. (2023), however, rigorous experimental tests of theoretical results on mediation are lacking.[6]

Beyond the specific focus on mediation, our work tests the ability of experimental participants to use sophisticated strategies to convey and extract information in the lab. It recalls recent experimental studies on Bayesian persuasion (Frechette et al. 2022; Nguyen 2017; Au and Li 2018; Aristidou et al. 2019), in which one group of experimental participants designs the information structure and another group reacts to the signals, which may obfuscate the payoff-relevant state. These studies find that, while participants react to information, behavior departs significantly from the rational benchmark. In our experiment, the mediation mechanism is not chosen by participants, but is fixed and commonly known. And yet, the problem remains very rich because the payoff-relevant state (the agents' types) must be elicited, giving rise to truth-telling constraints that must be satisfied in the optimal equilibrium.

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. One concern that appears in the literature is whether or not real-life mediators can commit to a mediation protocol as required in the theory, and in particular commit to not recommending a peaceful division under some circumstances,

[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.

[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), and Roth (2016). For VCG mechanisms for public good provision, see for example Attiyeh et al. (2000), Chen and Plott (1996), and Chen (2008).

de facto triggering conflict.[8] 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. Our work cuts through this debate by showing that, even with commitment power, mediation falls short.

The paper proceeds as follows. The next section describes the model and its main theoretical properties, comparing optimal mediation and direct communication; Section 3 discusses the experimental hypotheses; Section 4 describes the experimental design; Section 5 reports the results; Section 6 examines possible reasons why the optimal mediation algorithm is not more successful than direct communication in reaching agreement in the lab; Sections 7 and 8 discuss additional evidence collected in later robustness sessions: Section 7 shows that results remain unchanged using a mediation mechanism with strict incentive constraints; Section 8 presents some evidence on the frequency of peaceful resolution in the absence of communication; finally, Section 9 concludes.

2 The model

The mediation game we took to the lab follows closely the model in Hörner, Morelli and Squintani (2015), referred to as HMS in what follows. 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, a dispute follows, 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 agreement.

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 $1/2 \geq (1-q)\theta + q\theta/2$, and thus peace can be easily implemented. To highlight the role of information, HMS (and we) focus on the more interesting case where $1/2 < (1-q)\theta + q\theta/2$, or $q < (2\theta - 1)/\theta$.

[8] The mediator's role in triggering conflict is one uneasy implication of the theory. Note however that the mediator's only role is to manage information by issuing non-binding recommendations. Commitment to "triggering conflict" under some conditions stands for revealing information such that conflict is the outcome of the subgame. This can be thought of as refusing to mediate, or "walking out" as we phrase it in the lab.

The core of the analysis is the procedure through which the two players can reach an agreement. We consider two alternative procedures: direct communication and mediated communication. In both cases, the players take actions in two consecutive stages: a message stage and an allocation stage.

Under direct communication, after learning one's own type, at the message stage each player sends a cheap talk message $m(T)$ to the other player. 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. The demand may consist of the refusal to negotiate, or indicate the demanded share of the resource. With an eye to the experimental implementation, 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 agreement prevails. If either player chooses w, or if $d_1 + d_2 > 1$, then no agreement is reached and conflict follows: the resource shrinks to θ and is divided according to the players' types.[9] 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.

Under mediated communication, a third party enters the game, the mediator, whose objective is to maximize the probability of agreement. The mediator shares the common prior q but has no information on the realizations of the players' types and no power to enforce any recommendation. 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. The mediator commits to a recommendation $r(m_1, m_2)$ which we constrain to one of the following values $r \in \{(1-\theta, \theta), (1/2, 1/2), (\theta, 1-\theta), w\}$. As before w stands for "walking out", or the mediator's refusal to mediate. If the mediator makes a recommendation, then each player has the option to either obey and accept the recommendation, or deviate by 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 disagreement follows, the resource shrinks to θ and is divided according to the players' types.[10]

[9] 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.

[10] In the HMS protocol and in our lab implementation, war is directly triggered following the failure of mediation. If another round of direct communication were allowed to take place, war would be an equilibrium outcome of all following

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

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

The result follows if the frequency of H's is neither too high nor too low. It cannot be too high because, as we mentioned above, for $q > (2\theta - 1)/\theta$ there always exists an uninformative equilibrium under direct communication in which every type accepts 1/2, and the probability of peace is 1. But the frequency of H's should also not be too low, and the intuition here is more subtle. The essential difference between mediated communication and direct communication is that messages to the mediator are confidential. Confidentiality allows the mediator to issue recommendations that reveal the messages only partially. More precisely, the mediator may be able to induce H types to accept 1/2, even though the opponent is an L with positive probability. However, when L types are too common, the mediator is unable to both offer the equal split to (H, L) with positive probability and convince H to accept it.

To better understand the result, consider the following. By the revelation principle (Myerson, 1982), there is an optimal mediation program that is also a direct revelation mechanism. Under such optimal 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 obedience constraint (H's incentive to accept 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 conflict when matched with an L and increasing peace)—although all messages are sincere, the opponent's type is *obfuscated*.

subgames. Both HMS' model and our experiment select such an equilibrium in the event of continued bargaining. In the lab, this also has the virtue of simplifying the mechanism for subjects.

Strictly speaking, the mechanism we analyze and play in the lab is not a direct revelation mechanism because we allow players to remain silent by sending message s instead of reporting a type (h or l). In such a case, we assume the mediator interprets s as h with probability q and as l with probability $1-q$ (i.e., according to the prior). In so doing, we deviate slightly from HMS. However, this has no substantive implications for the theory: every equilibrium of the mechanism in which silence is never used is one-to-one with an equivalent equilibrium of the direct revelation mechanism.[11] In particular, the truthful-obedient equilibrium of our mechanism remains optimal among all possible mechanisms. We include silence for experimental reasons only, to give inexperienced subjects the intuitive option of hiding their type while they learn the game, without complicating the data with random exploratory messages.

HMS compare the optimal mediation mechanism to the best equilibrium of the direct communication game. In the direct communication game, they allow the players to have access to a public correlation device or a "peace conference." After exchanging messages, the device publicly recommends (stochastically as a function of the messages) a split of the resource, which can then be accepted or rejected by each party, independently. The equilibrium corresponds to the optimal correlated equilibrium with public messages and public signals. In such an equilibrium, messages are sincere and because they are public, the recommendation corresponds to the outcome of the optimal mediation program without obfuscation. Thus, obfuscation is central to Proposition HMS: if obfuscation cannot be supported, the optimal mediation program cannot deliver a strictly higher probability of peace than the best equilibrium of the direct communication game with public messages and public signals.[12]

In the lab, however, there is no public correlation device. In its absence, achieving the optimal correlated equilibrium of the direct communication game through the randomization of individual messages is in practice impossible.[13] As a result, any equilibrium of the direct communication game played in the lab must result in a weakly lower frequency of peace than the upper bound established in HMS. It then follows that the frequency of peace must be weakly lower than in the optimal mediation equilibrium, and strictly lower if $(2\theta-1) < q < (2\theta-1)/\theta$.

The public correlation device exploited in the best equilibrium of the direct communication game

[11] In mechanisms without obfuscation, the mediator's recommendation reveals to the silent player how her s message was interpreted, and hence silence is nothing more than a randomization device. In mechanisms with obfuscation, the mediator's recommendation need not reveal how the message was interpreted, in which case the player remaining silent loses some information.

[12] As noted in HMS, the public correlation device can be replicated by an additional round of communication between the players using jointly controlled lotteries (Aumann and Hart 2003; Krishna 2007; Krishna and Morgan 2004), as opposed to public correlation of play.

[13] Mixed strategy profiles cannot typically result in correlated randomness (Forges 1986).

induces war with a positive probability in response to specific pairs of messages. The possibility of war plays a disciplining role in equilibrium that mimics the commitment demanded from the mediator, and in particular induces truthful messages. In the lab, absent the public correlation device, not only is the equilibrium frequency of conflict higher, but messages also fail to be truthful.

In studying equilibria under direct communication, and in fact in all protocols we consider throughout the paper, we concentrate on Perfect Bayesian Equilibria (PBE) in undominated strategies that are symmetric for players of a given type. We can state:

Proposition 1. *Suppose $\theta/2 > 1-\theta$, and consider the direct communication game played in the lab. In any symmetric PBE in undominated strategies, at least one type of player must be lying with strictly positive probability.*

Proof. Recall that $\theta/2 > 1-\theta$ is one of the model's maintained assumptions. It implies that θ is high enough that fighting an opponent of the same type yields more than accepting the lower share $(1-\theta)$. In addition, note that $d = w$ (walking out) is weakly dominated by $d = \theta$ (making a large demand). Thus in any PBE in undominated strategies, $d = w$ is never played.

Suppose then, contrary to Proposition 1, 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 dominates demanding $1-\theta$.[14] 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 herself 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 his 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 the 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. □

Proposition HMS and Proposition 1 establish the qualitative hypotheses at the heart of our study: the best equilibrium under mediated communication can yield both higher peace and higher sincerity

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

than can be achieved in the direct communication game played in the lab. In the next section, we derive precise numerical predictions for the frequency of truthfulness and peace under the parameter values used in the experiment. As typical of communication games, this exercise requires selecting over a large set of equilibria, even when restricting attention to symmetric PBEs in undominated strategies. Proposition HMS and Proposition 1, holding over a wider range of equilibria, are the stronger foundations for the experimental aims of this project.

3 Experimental parametrization and predictions

In taking the HMS model to the lab, we made two modifications, neither of which affects the model's theoretical properties. First, as described earlier, we allow players to send a silent message.[15] Second, we constrain both demands and the mediator's recommendations to lie in a restricted set. The set includes all values that can appear in equilibrium, thus simplifying the subjects' problem without affecting equilibrium predictions.

Throughout the experiment we fixed $\theta = 0.7$, ensuring that Proposition 1 applies. We studied two different parametrizations of the ex-ante frequency of H types: $q = 1/2$ and $q = 1/3$. Proposition HMS applies to $q = 1/2$, but not to $q = 1/3$.

3.1 Mediated communication

In the mediated communication (MC) treatment, the optimal program is implemented as a computer algorithm, responding to subjects' messages. The program follows directly from Lemma 3 in HMS and issues the following recommendations:[16]

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

$\underline{q = 1/3}$. $r(l,l) = (0.5, 0.5)$; $r(h,l) = \{(0.7, 0.3)$ *with probability* $3/4$, w *otherwise*$\}$; $r(h,h) = w$.

In the best equilibrium of this program, messages are fully sincere and silence is not used, and if the mediator makes a recommendation, the recommendation is accepted by both players with probability 1, regardless of type.

[15] Later auxiliary sessions without silent messages yield closely comparable results. We discuss them in the online appendix (Section B.5.4).

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

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 agreement under optimal mediation is lower at lower q: as can be readily calculated, the expected frequency of peace in the best equilibrium is $\frac{7}{8}$ if $q = 1/2$, and $\frac{7}{9}$ if $q = 1/3$.

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 prefers a dispute when matched to 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.[17] With $q = 1/3$, on the other hand, $q < 2\theta - 1$, and the optimal 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 parametrizations interesting. Theory tells us that it is the possibility of obfuscation that renders the mediator indispensable; but obfuscation also complicates the subjects' problem. Collecting data under both $q = 1/2$ and $q = 1/3$ allows us to study how subjects react to optimal mediation programs with and without obfuscation.

3.2 Direct communication

The direct communication (DC) treatment is a Nash demand game with cheap talk messages. It admits a large number of equilibria, even when restricting attention to PBEs in undominated strategies. Here we discuss equilibrium sets with two desirable properties. First, demand strategies are either conditioned on type only, or on type and a single set of messages (as opposed to both messages sent and received). Thus the equilibria are relatively simple, an important asset for a lab experiment. Second, although each equilibrium set is large, the probability of peace—the outcome that most matters to us—is constant across the whole set. Proposition A1 in the appendix (Section A.1) characterizes the equilibria below for arbitrary q and θ. Here we report equilibrium demands and messages when specialized to the experimental parameters.

For $q = 1/3$—that is, when the frequency of H types is relatively low—a particularly intuitive class

[17] 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 best equilibrium.

of equilibria exists in which demand strategies are pure and do not depend on messages: H types always demand 0.7, and L types always demand 0.5. Note that such demands cannot be best responses to an H opponent. Hence equilibrium requires that, following messages, the posterior probability of the opponent being H must be low enough, constraining the range of acceptable message strategies. As long as such constraints are satisfied, however, messages are irrelevant to demands, and mixing over messages is indeed a best response at the message stage.

We denote by τ_T the probability that type T's message is truthful, and by σ_T the probability that T sends a silent message. Calling $\delta_d(T, m, m')$ the probability that type T who sent message m and received message m' demands d, we find:[18]

Equilibria 1: $q = 1/3$. *At the demand stage:* $\delta_{0.7}(H, m, m') = 1$, $\delta_{0.5}(L, m, m') = 1$ *for all* m, m'. *At the message stage, for any* $(\tau_L + \sigma_L) \in (0, 1)$*:*

$$\tau_H \in [\max\{0, 1 - \sigma_H - (8/3)\tau_L\}, \min\{1, (8/3)(1 - \tau_L - \sigma_L)\}]$$
$$\sigma_H \leq (8/3)\sigma_L.$$

Across this set of equilibria, the L type can be arbitrarily close to truthful, but the more truthful the L type, the less truthful the H type can be. In line with Proposition 1, there cannot be full truthfulness, the one prediction that emerges sharply from the constraints on messages. The equilibrium demand strategies are instead uniquely pinned down.

Such simple equilibria, with deterministic actions, do not exist when the frequency of H types is higher (when $q = 1/2$). A different set of equilibria exists, in which L types mix over different demands. These equilibria, which we denote as equilibria 2, can be sustained for either of the frequencies of H types we consider in the experiment (that is, whether $q = 1/2$ or $q = 1/3$.)

Equilibria 2: $q = 1/2$ and $q = 1/3$. *At the demand stage:* $\delta_{0.7}(H, m, m') = 1$ *for all* m, m'*;* $\delta_{0.7}(L, m, m') = 2\left[1 - (3/7)\left(\frac{1}{1-\pi_{m'}(q)}\right)\right] = 1 - \delta_{0.3}(L, m, m')$ *for all* m'*, where* $\pi_{m'}(q)$ *is the posterior probability the opponent is* H *given message* m'*. At the message stage, both types randomize. For any* $(\tau_L + \sigma_L) \in (0, 1)$*:*

[18] The constraints on H's messaging strategy appear cumbersome, but in fact amount simply to $\pi_{m'} \leq 2\theta - 1$, where $\pi_{m'}$ is the posterior probability the opponent's type is H after having received message m', with $m' \in \{l, h, s\}$. For example:

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

If $q = \frac{1}{2}$:

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

$$\sigma_H \in [(1/6)\sigma_L, (4/3)\sigma_L].$$

If $q = \frac{1}{3}$:

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

$$\sigma_H \in [(1/3)\sigma_L, (8/3)\sigma_L].$$

In this second set of equilibria, again H types demand 0.7 with probability 1, and again both types are indifferent over messages, but now the frequency of H types (q) is too high for L types to always demand 0.5 and risk conflict. They compromise by randomizing between playing safe and demanding 0.3, and mimicking H's and demanding 0.7.

In both sets of equilibria, the message probabilities cover a large range, and in particular always include the possibility of non-informative messages ($\sigma_H = \sigma_L$ and $\tau_H = 1 - \tau_L - \sigma_L$). Partially informative messages—messages more likely to be sent by one type rather than the other—are possible too but, for a given set of equilibria (Equilibria 1 or Equilibria 2), the message strategies have no impact on the ex ante probability of peace. In the case of Equilibria 1, peace can occur only when the two opponents are both L types, an event which occurs with probability $(1-q)^2$, or $\left(\frac{2}{3}\right)^2 = 0.444$. In the case of Equilibria 2, peace occurs if and only if at least one of the two players is an L who demands 0.3. Demand probabilities vary with the message sent but, as we show in more detail in the appendix, for given q, the unconditional probability of an L type demanding 0.3 is constant over the whole set of equilibria, and as a result, so is the ex ante probability of peace. Even when the message is informative, the mixing probabilities at the demand stage effectively nullify the information provided by the message. Semi-pooling equilibria where types partially distinguish themselves through their messages do not have higher peace than when communication is fully uninformative.

3.3 Comparative theoretical predictions for MC and DC

Table 1 reports numerical predictions for the expected frequency of peace (P) and of truthful messages in the two treatments, under the equilibria characterized above.[19]

[19] In the DC equilibria, the messaging for L is unconstrained and pins down the range of messaging strategies for H. The ranges for H messages given in the table are based on the empirical messaging strategies for L observed in the experiment.

$q = 1/2$						
		Peace	H's messages		L's messages	
MC		P	τ_H	σ_H	τ_L	σ_L
	Equil:	0.875	1	0	1	0
DC		P	τ_H	σ_H	$\tau_L + \sigma_L$	
	Equil 2:	0.586	[0.75,0.79]	[0.02,0.13]	(0,1)	

$q = 1/3$						
		Peace	H's messages		L's messages	
MC		P	τ_H	σ_H	τ_L	σ_L
	Equil:	0.778	1	0	1	0
DC		P	τ_H	σ_H	$\tau_L + \sigma_L$	
	Equil 1:	0.444	[0.1,1]	[0,0.40]	(0,1)	
	Equil 2:	0.345	[0.70,0.73]	[0.05,0.40]	(0,1)	

Table 1: Predictions: MC and DC. P is the expected frequency of peace; τ_T is the probability of a truthful message by type T; σ_T is the probability of a silent message by type T.

In both the DC and MC treatments, the probability of peace is predicted to be lower under $q = 1/3$ than under $q = 1/2$ (in the case of DC, this is true regardless of which of the two equilibria are played under $q = 1/3$). When the frequency of H types is smaller, L types act more aggressively, and the result is more frequent conflict.

The DC equilibria we characterize are intuitively plausible, given their simplicity, and provide a concrete reference for the experiment. We know, however, that other equilibria exist—equilibria where demand strategies respond both to the messages sent and to the messages received. Thus, although we will refer to Table 1 when describing the experimental results, our stronger theoretical foundation is Proposition HMS: *any* equilibrium of the DC game must deliver a frequency of peace that is lower than the best equilibrium of the MC game (and strictly so if $q = 1/2$).

4 Experimental design

We ran the experiment at Columbia's Experimental Lab for the Social Sciences (CELSS) with subjects recruited through the lab's ORSEE recruitment system (Greiner, 2015). Most subjects were undergraduate students at Columbia University and Barnard College. The experiment lasted about 90 minutes and earnings ranged from \$16 to \$37, with an average of \$28 (including a \$10 show-up fee). Experimental procedures were standard and are described in detail in the online appendix (Section B.6), where the instructions for one of the treatments are reproduced.[20]

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

Subjects in each experimental session were exposed to a single parametrization, either $q = 1/2$ or $q = 1/3$, and to both the direct communication and mediated communication treatments. Each treatment consisted of multiple rounds, and instructions for each part were read just before that part began. 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.

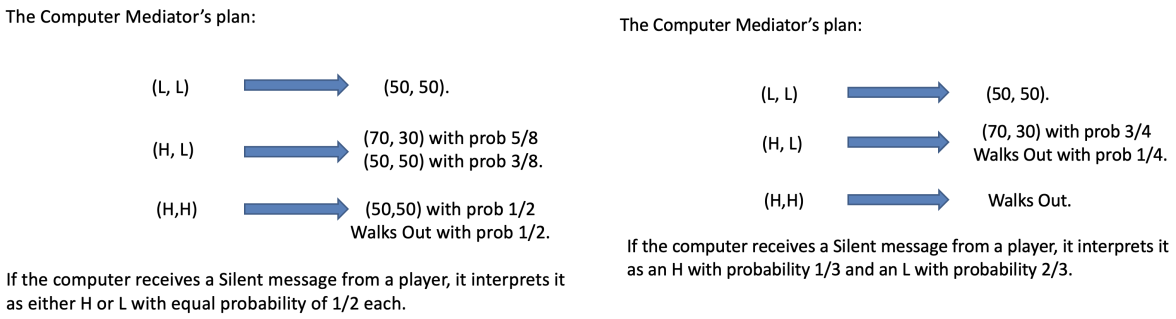
Direct communication (DC). The DC treatment corresponds exactly to the direct communication game described in the previous section. After being randomly matched in pairs and assigned a type according to q, all subjects simultaneously sent their partner a message, chosen among $\{h, l, s\}$. After messages were exchanged and read, each player, again simultaneously, 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 randomly. In each session, subjects played 20 rounds of the DC treatment.

Mediated communication (MC). In the mediated communication treatment, we introduced the mediator, delegating the mediator's role to the computer, programmed to implement the optimal mediation program. After having been randomly matched in pairs and privately assigned types, each subject sent to the computer-mediator a confidential 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 parametrization 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 $1 - q$.[21]

Unless the computer chose w, the recommendation was conveyed to each subject who then chose, separately, either to accept it or reject it. The 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 learned their payoff from the round. Note that if the computer-

[21] To avoid confusion, in the instructions, we used the same notation (upper-case letters) for both types and messages. We communicated clearly that messages did not need to be truthful.

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 MC treatment.



(a) q=1/2 (b) q=1/3

Figure 1: The Mediation program as shown to subjects

For each parametrization, $q = 1/2$ or $q = 1/3$, we ran six experimental sessions, each with 12 subjects and both treatments. We varied the order of treatments so as to treat DC and MC symmetrically: for a given value of q, three sessions had order DC, MC, and three had order MC, DC. The design also allows us to compare DC and MC between subjects, when each treatment is run first.[22]

Because the game is challenging, we began all sessions with 10 rounds of a simplified scenario where subjects exchanged demands without a prior message stage. These rounds helped subjects understand the interface, the probabilistic nature of the types, and the cost of disagreement, without the added strategic problem of sending and interpreting messages and recommendations. They were designed for training but, given their length, we chose to reward them. We return to these introductory rounds in Section 8, where we discuss them alongside auxiliary sessions run after the conclusion of the original experiment.

In what follows, we use the term "treatment" to refer to the two negotiation procedures—DC or MC—and the term "parametrization" to refer to the two possible values of parameter q. All data are deposited in OSF at https://osf.io/tkaw2/.

	Peace	H's messages		L's messages		H's acceptance		L's acceptance	H's payoff	L's payoff
	P	τ_H	σ_H	τ_L	σ_L	α_h	α_l	β_l	pay_H	pay_L
$q = 1/2$										
Equil:	0.875	1	0	1	0	1	—	1	0.525	0.438
Data:	0.547	0.87	0.04	0.57	0.11	0.63	0.17	0.92	0.509	0.353
$q = 1/3$										
Equil:	0.778	1	0	1	0	—	—	1	0.583	0.408
Data:	0.464	0.66	0.09	0.63	0.12	—	0.19	0.82	0.562	0.344

Table 2: Data and predictions: MC

$q = 1/2$									
	Peace	H's messages		L's messages		H's payoff	L's payoff		
	P	τ_H	σ_H	τ_L	σ_L	pay_H	pay_L		
Equil 2:	0.59	[0.75,0.79]	[0.02,0.13]	(0, 1)	(0, 1)	0.525	0.300		
Data:	0.57	0.81	0.12	0.31	0.10	0.511	0.346		
	H's demands			Ll's demands			Lh's demands		
	$\delta_{70}(H)$	$\delta_{50}(H)$	$\delta_{30}(H)$	$\delta_{70}(Ll)$	$\delta_{50}(Ll)$	$\delta_{30}(Ll)$	$\delta_{70}(Lh)$	$\delta_{50}(Lh)$	$\delta_{30}(Lh)$
Equil 2:	1	0	0	[0.49,0.92]	0	[0.08,0.51]	[0, 0.06]	0	[0.94, 1]
Data:	0.52	0.40	0	0.01	0.41	0.56	0.14	0.79	0.07
$q = 1/3$									
	Peace	H's messages		L's messages		H's payoff	L's payoff		
	P	τ_H	σ_H	τ_L	σ_L	pay_H	pay_L		
Equil 1:	0.44	[0.10, 1]	[0, 0.40]	(0,1)	(0,1)	0.583	0.333		
Data:	0.49	0.74	0.10	0.30	0.15	0.568	0.334		
	H's demands			Ll's demands			Lh's demands		
	$\delta_{70}(H)$	$\delta_{50}(H)$	$\delta_{30}(H)$	$\delta_{70}(Ll)$	$\delta_{50}(Ll)$	$\delta_{30}(Ll)$	$\delta_{70}(Lh)$	$\delta_{50}(Lh)$	$\delta_{30}(Lh)$
Equil 1:	1	0	0	0	1	0	0	1	0
Data:	0.71	0.17	0	0.07	0.52	0.38	0.25	0.59	0.16

Table 3: Data and characterized equilibrium strategies: DC.

5 Experimental results

Tables 2 and 3 compare theoretical predictions for all strategies and the ex-ante probability of peace to their empirical counterparts from the lab, for each parametrization and both treatments. Because our focus is the comparative performance of the two treatments, most of the analysis will center on the two treatments' frequency of truthfulness for each type (τ_H and τ_L) and the frequency of peace, P, data that are directly comparable across DC and MC.

Before doing so, however, we make two comments on demand strategies in the DC game. First,

[22] With both orders, in-between DC and MC, we ran an exploratory treatment where the mediator was played by an experimental subject without a pre-specified mediation program. We discuss theoretical and experimental results in Casella, Friedman, and Perez Archila (2020), but omit them here. Robustness sessions reported in the online appendix (Section B.5.2) show that the comparison between DC and MC is not affected by the presence of such a treatment (see also footnote 50 below).

having characterized two sets of equilibria for $q = 1/3$, we find that one is clearly closer to the data. What we called equilibrium 1 matches the high propensity of subjects to demand 70 if H and 50 if L. In Table 3 and in what follows, we use it as our theoretical reference for data under DC and $q = 1/3$.[23]

However, such an equilibrium does not exist for $q = 1/2$. With $q = 1/2$, the set of equilibria we identify requires H types to always demand 70, an action we observe with only 50 percent frequency, and requires L types never to demand 50, something we instead observe with almost 66 percent frequency. Thus, and this is the second comment, the equilibria we characterized do not explain the demands we see in the lab when $q = 1/2$.[24] Once again, the possibility to rely on Proposition HMS is valuable: the superiority of the best equilibrium of the HMS mechanism does not depend on equilibrium selection in the DC game.

Tables 2 and 3 also report average payoffs for H and L types, in the theory and in the data. According to the theory, the higher probability of peace under MC translates into higher expected payoffs for the L types only: the H types' payoffs are predicted to be identical across MC and DC. In the data, noise predictably reduces payoffs for both types and under both communication protocols, but the loss is particularly noticeable for L's under mediation: the higher frequency of conflict, relative to the theory, translates into larger losses for the weaker type.

We begin our comparative discussion of MC and DC by looking at messages.

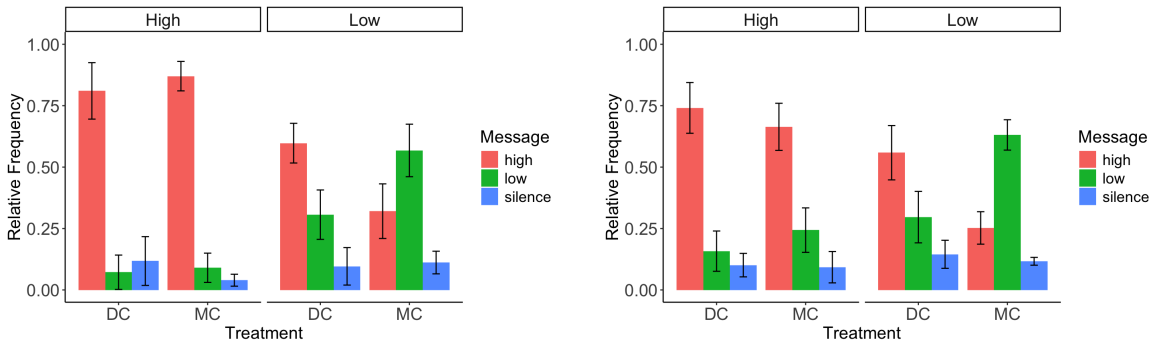
5.1 Sincerity

The two panels of Figure 2 report the frequencies of different messages in the two parametrizations, $q = 1/2$ and $q = 1/3$, for both treatments. 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, as in all subsequent tests and statistics.

The figure makes clear a number of regularities. First, although we never see full sincerity, H types send message h with high frequency in both treatments and for both parametrizations. Whether under DC or MC, more than 80 percent of all H types send message h when $q = 1/2$; more than 65 percent do so when $q = 1/3$. H types' sincerity increases from DC to MC, if $q = 1/2$, while it decreases

[23] As in Table 1, the numerical values reported as predictions for DC in Table 3 are based on the observed frequencies of L's messages, which the theory leaves unconstrained and which anchor the remaining strategies.

[24] Because silent messages are few and Table 3 is already dense, we did not include demands conditioned on silent messages. The evidence however is consistent: following silence, most L's demand 50, in line with equilibrium 1 and contradicting equilibrium 2.



(a) $q = 1/2$ (b) $q = 1/3$

Figure 2: Messages by type and treatment

if $q = 1/3$, but the differences between treatments are not statistically significant at either value of q. Note that H types' truthfulness should not be taken for granted: even under MC, the incentive compatibility constraints of the H types are binding when taking into account the possibility of double deviation (an untruthful message followed by rejection of the mediator's recommendation).

Second, there is less sincerity but a clear treatment effect for L types: the frequency of l messages from L subjects goes from 31 percent in DC to 57 percent in MC 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 prediction of full sincerity implied by the best equilibrium under MC is not observed in the data, but for L types, MC strongly increases sincerity relative to DC.

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 4, a linear probability model confirms what the figures show.[25] L types are less sincere than H types, and for L types treatment effects are present and significant: sincerity is lower under DC than under MC. 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.[26]

Finally, as shown in the second two columns of Table 4, the use of silence is not only scarce but declines with experience. Although an intuitive choice, silence is a redundant option, and one that

[25] 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).

[26] 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 online appendix (Section B.5.1), we report Kullback Leibler (KL) measures applied to our data. It remains true that the treatment conveying most information is MC, for both messages and in both parametrizations.

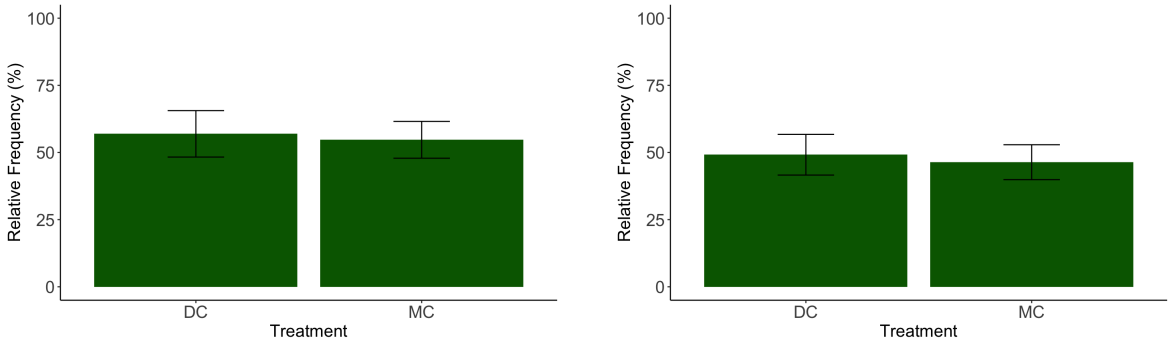
	Dependent variable:			
	Sincerity		Silence	
	$q = 1/2$	$q = 1/3$	$q = 1/2$	$q = 1/3$
MC treatment	0.061	−0.077	−0.079**	−0.008
	(0.062)	(0.063)	(0.034)	(0.020)
Second treatment	0.073	0.140**	−0.089***	−0.043**
	(0.061)	(0.064)	(0.034)	(0.021)
L-type	−0.596***	−0.388***	−0.074	0.035
	(0.089)	(0.145)	(0.065)	(0.029)
Round	−0.003	−0.003	−0.0002	−0.001
	(0.002)	(0.004)	(0.001)	(0.002)
MC treatment × L-type	0.198***	0.412***	0.096**	−0.020
	(0.076)	(0.103)	(0.043)	(0.040)
Second treatment × L-type	0.093	−0.107	0.036	0.030
	(0.075)	(0.105)	(0.043)	(0.040)
Round × L-type	0.004**	−0.001	0.003***	−0.0005
	(0.002)	(0.003)	(0.001)	(0.002)
Constant	0.802***	0.702***	0.164***	0.129***
	(0.066)	(0.086)	(0.050)	(0.034)
Observations	2,880	2,880	2,880	2,880
Note:				*p<0.1; **p<0.05; ***p<0.01

The excluded (default) categories in the regression are the H-Type, the DC treatment, and the first of the two treatments in the session (DC or MC). Round refers to the round number within the treatment. Standard errors are clustered at the session level.

Table 4: Sincerity and Silence.

should not be used in the optimal equilibria. Its decline with experience is a useful check on subjects' attention and understanding of the game.

5.2 Peace



(a) $q = 1/2$ (b) $q = 1/3$

Figure 3: Frequency of peace

Figure 3 reports the frequency of peace across the two treatments, for both parametrizations, as well as 95 percent confidence intervals (with standard errors clustered at the session level). Ordering the numbers as $\{DC, MC\}$, the frequencies of peace in the data are: $\{0.57, 0.55\}$ if $q = 1/2$, and $\{0.49, 0.46\}$ if $q = 1/3$. In both cases, DC results in slightly more frequent peace, but the effect is very small. Whether $q = 1/2$ or $q = 1/3$, there is no significant difference between DC and MC. The highest theoretical frequency under MC (87 percent with $q = 1/2$, and 78 percent with $q = 1/3$) is (very far) outside the confidence intervals. The prediction for our class of equilibria under DC (57 percent with $q = 1/2$ and 44 percent with $q = 1/3$) is instead broadly consistent with the data.

On the other hand, the difference in peace between the two parametrizations is in the direction the theory predicts, with higher peace in both treatments under $q = 1/2$, a finding we study in more detail and confirm in the online appendix (Section B.5.3).

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 5, 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 slightly higher when the treatment is played later in the session, and for $q = 1/2$ it decreases slightly though significantly over time. Across treatments, DC and MC, peace is closely comparable. The conclusion

	Dependent variable:			
	Peace			
	$q = 1/2$	$q = 1/3$	$q = 1/2$	$q = 1/3$
MC treatment	-0.022	-0.028	-0.017	-0.017
	(0.051)	(0.040)	(0.048)	(0.023)
Second treatment	0.072	0.081^{**}	0.059	0.063^{***}
	(0.051)	(0.040)	(0.049)	(0.023)
Round	-0.004^{***}	-0.0003	-0.002^{**}	0.001
	(0.001)	(0.003)	(0.001)	(0.003)
Pair type *H-L*			0.360^{***}	0.313^{***}
			(0.037)	(0.026)
Pair type *L-L*			0.686^{***}	0.723^{***}
			(0.052)	(0.031)
Constant	0.572^{***}	0.454^{***}	0.213^{***}	-0.014
	(0.056)	(0.057)	(0.069)	(0.051)
Observations	1,440	1,440	1,440	1,440
Note:				*p<0.1; **p<0.05; ***p<0.01

The excluded (default) category is the DC treatment, and the first of the two treatments in the session (DC or MC). When looking at different pair types, the default pair is *H-H*. Round refers to the round number within the treatment. Standard errors are clustered at the session level.

Table 5: Peace.

remains qualitatively unchanged if we study the frequency of peace in MC and DC between subjects, comparing sessions with different orders of treatments, such that each treatment is played first in the session (see Section B.5.2 in the online appendix).[27]

6 Mediation increases sincerity but not peace. Why?

The lesson from the data is unambiguous: because of the messages by L types, sincerity is higher and messages more informative under MC. But peace is not. We can represent both observations in a single graph.

Recall that τ_T denotes the frequency with which a type T player sends a truthful message, and σ_T the frequency with which the player is silent. Recall also that in MC the computer-mediator 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 MC and DC, for the purposes of this figure only, we also code silent messages in DC according to the prior (i.e., as h with frequency q).[28]

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 DC treatments, and red spheres for MC treatments. The two green cubes correspond to the theoretical equilibria with highest peace in the two treatments (the green cube centered among the yellow spheres refers to DC; the green cube corresponding to $\widehat{\tau}_L = 1$ and $\widehat{\tau}_H = 1$ represents the HMS equilibrium in MC).[29]

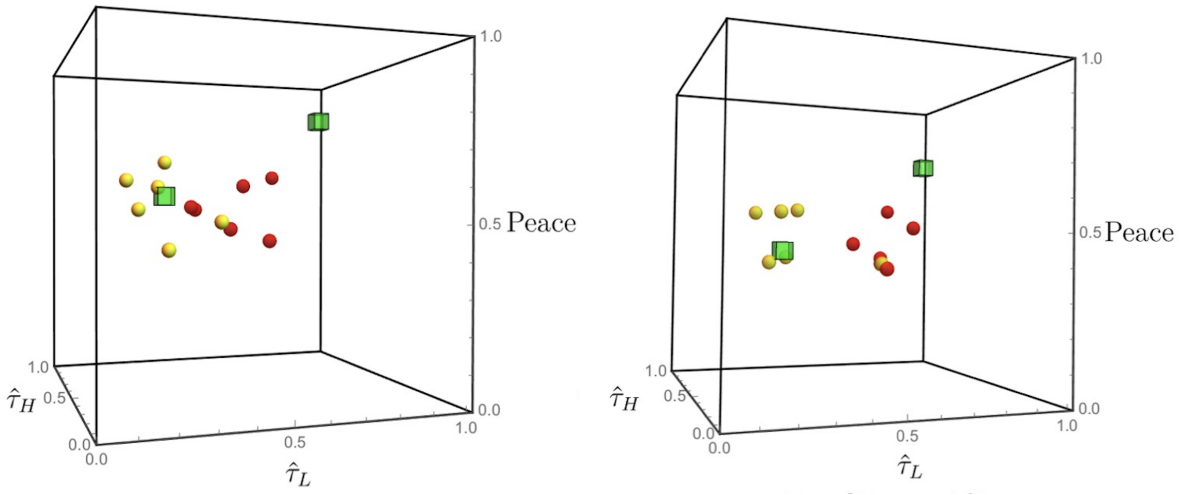
As shown earlier, the two treatments on average yield similar values for $\widehat{\tau}_H$, but different values for $\widehat{\tau}_L$. Here, yellow and red spheres align similarly along the depth axis, but are clearly differentiated along the horizontal axis, and the orientation of the figures highlights the two clusters, almost fully distinct, with lower $\widehat{\tau}_L$ values for DC, and higher $\widehat{\tau}_L$ values for MC. 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[30]

[27] Section B.5.3 in the same appendix reports regressions with a full set of interaction terms. The conclusion remains the same.

[28] Silence is rarely used, and we have verified that the plot hardly changes under any other possible imputation.

[29] For DC, the green cube corresponds to the equilibrium closest to the data among the equilibria characterized earlier.

[30] To better visualize the data, the reader may view animations of Figure 4 and other 3D figures that appear later at https://www.youtube.com/watch?v=8NIeQ1N_KNo.



(a) $q = 1/2$ (b) $q = 1/3$

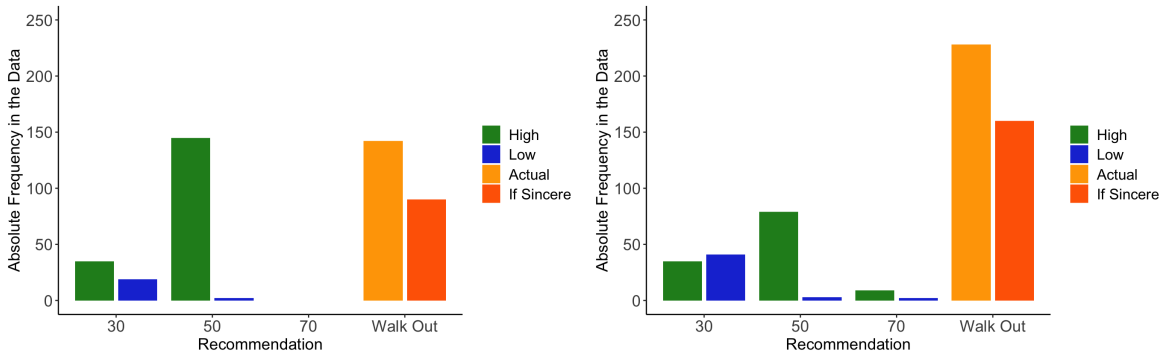
Figure 4: Sincerity and peace in DC and MC sessions. For each session, yellow spheres report the data from DC rounds and red spheres report data from MC rounds, both aggregated at the session level. The green cube at lower $\widehat{\tau}_L$ values is the equilibrium prediction for DC; the green cube at $\widehat{\tau}_L = \widehat{\tau}_H = 1$ is the optimal HMS equilibrium for MC.

Why was the promise of optimal mediation not realized in the lab setting? Figure 5 gives some indications of where the problems lie. The figure plots, for each parametrization, the causes of war under MC in the data. The orange columns correspond to the mediator'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 mediator's offer by H types, and blue columns by L types, organized according to the offer.[31] In the optimal equilibrium, all messages are sincere, all recommendations are accepted, and conflict 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 parametrizations, 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 by L types, reflected in the higher frequency of refusals to mediate. The second, more striking, is the high number of rejections of proposed equal splits (recommendations of 50) by H types: sincere H types rejected more than one third of all $50 - 50$ splits they were recommended. The subtlety of the obfuscation does not appear 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

[31] The figures report individual rejections of the mediator's recommendation. Because a single rejection is sufficient to trigger conflict, there can be some double counting: two individual rejections can amount to a single recommendation being turned down.



(a) q=1/2 (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 the mediator's refusals to mediate, in the data and if players had been sincere respectively.

high frequency of refusals to mediate, and H's refusals of recommendations 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 MC treatment has multiple equilibria. Could it be that the best equilibrium under MC is fragile to the presence of lies?

6.1 Multiple equilibria and the fragility of peace

Keeping fixed the mediator's program, we study the equilibria of the MC treatment.[32] 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 (accounting for the option of silence).

[32] 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 $\theta/2 > (1 - \theta)$.

$q = 1/2$		$q = 1/3$	
(i)	$\alpha_h = 1, \hat{\tau}_L = 1, \hat{\tau}_H = 1$	(i)	$\hat{\tau}_L = 1, \hat{\tau}_H = 1$
(ii)	$\alpha_h = 0, \hat{\tau}_L = 1, \hat{\tau}_H = 1$	(ii)	$\alpha_l = 0, \hat{\tau}_L \in (0, 1)$,
(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)	$\alpha_l = 0, \alpha_h = 0, \hat{\tau}_L \in (0, 1)$,		
	$\hat{\tau}_H = 4/15 + (6/15)\hat{\tau}_L$		
(v)	$\alpha_l = 0, \alpha_h = 0, \hat{\tau}_L = 1, \hat{\tau}_H \in [2/3, 1)$		

Table 6: Equilibria under MC ($\beta = 1$)

We report the full set of equilibria in the appendix (Section A.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 time if $q = 1/2$, and 80 percent of the time if $q = 1/3$. In line with this observation, we focus here on equilibria with $\beta = 1$. The equilibria are reported in Table 6 and represented graphically in Figure 6.[33]

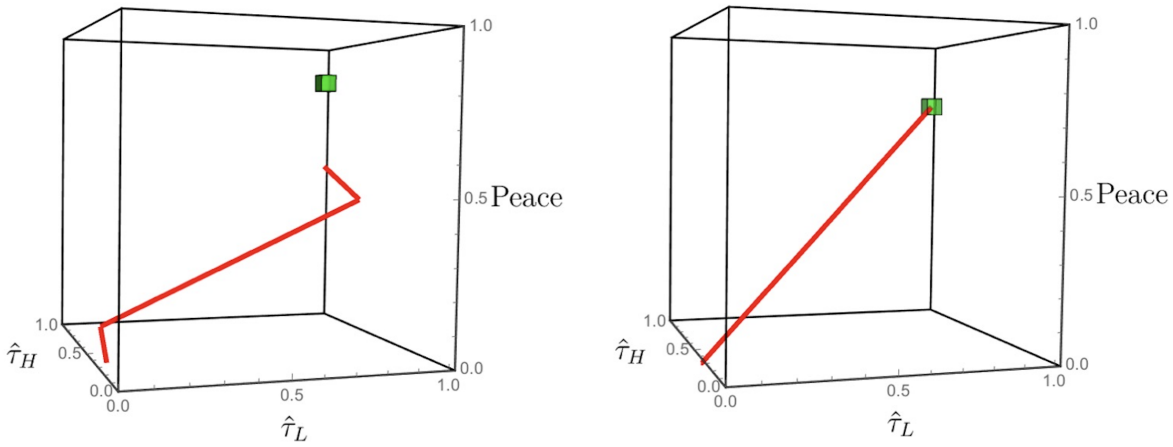
For both parametrizations, the first equilibrium in Table 6 is the HMS equilibrium, identified by the green cubes in Figure 6; the other equilibria correspond to the red lines. For both values of q, there are equilibria supporting a large range of peace probabilities, any frequency of truthfulness for L's, and almost as large a range for H's. Keeping fixed the mediator's program, equilibrium behavior under MC 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 2 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.[34] (We

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

[34] 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.



(a) $q = 1/2$ (b) $q = 1/3$

Figure 6: Equilibria under MC. The figure plots the equilibria in Table 6. The green cubes correspond to the HMS optimal equilibria.

include the possibility of silence in the appendix, specialized to the experimental parameters).

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

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

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

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

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

message m. Since all L types always accept $1/2$, it is not difficult to see that:

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

Using Bayes' rule:

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

and similar expressions for $Pr(Hl_j\ |(1/2,1/2),h_i)$ and $Pr(L_j|(1/2,1/2),h_i)$.[35] Using these and Equation (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 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 online appendix (Section B.2), 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

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

if higher probability is assigned to dominated rather than undominated actions. Under more plausible beliefs, convergence to the $\alpha_h = 1$ equilibrium is ruled out.[36]

Proposition 2 is very relevant for a lab experiment and possibly for actual applications of mediation plans, where some positive probability of lies 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.[37]

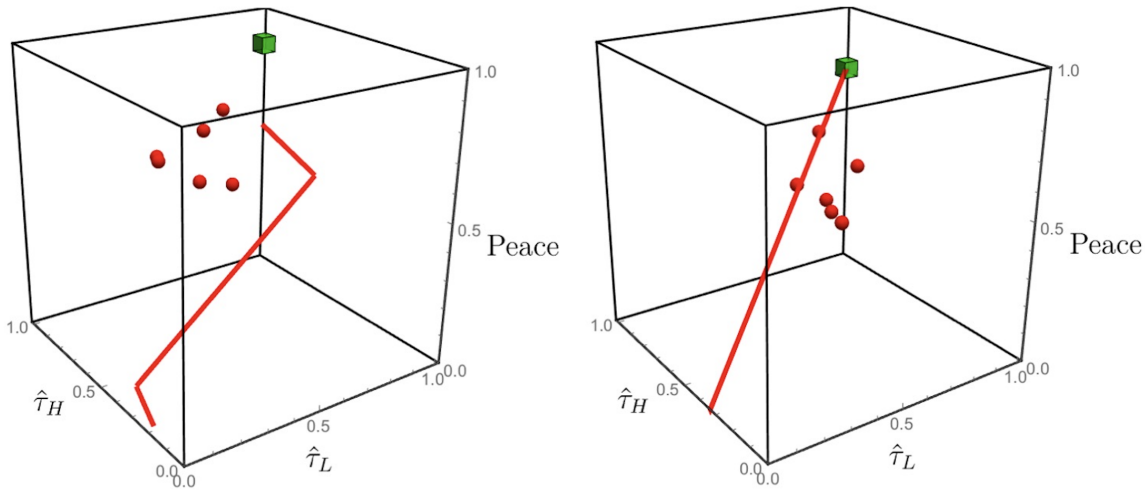
6.2 Sincerity and peace: Data v/s equilibrium predictions

Having observed the lack of predictive power of the HMS equilibrium, we ask if the data are 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.

As we already know, in both parametrizations and all sessions, sincerity, by either type, and peace, all fall short of the HMS equilibrium (the green cube). Figure 7 shows that, relative to the other equilibria, the deviations for the two parametrizations 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

[36] Note that because we are comparing actions across information sets, the argument cannot be used to claim that the equilibrium is not proper, in the sense of Myerson (1978).

[37] 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).



(a) q=1/2 (b) q=1/3

Figure 7: MC: Data and Equilibria. The green cubes and the red lines correspond to the equilibria in Table 6 (the green cubes correspond to the HMS optimal equilibria). The red spheres are the experimental data for MC, aggregated at the session level.

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.[38] But note that peace would have occurred only if the opponent sent message l[39] and accepted 30, that is, if the opponent was a sincere L. And in this case the payoff to the H type would be 70, whether from peace or from war. In other words, with $q = 1/3$, H's payoff from sincerity and compliance is identical to the payoff from message l (or s), and then the rejection of any recommendation of either 50 or 30. The loss of efficiency comes at no cost to the H player. In such a situation, it is 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 mediator, could explain the relatively high frequency of H's lies.

With $q = 1/2$, peace is instead higher than the equilibria predict. In all equilibria with less than perfect truthfulness H types never accept 50. In the data, aggregating over all sessions, the frequency of acceptances of 50 is just below 60 percent, with high dispersion across subjects. Why are H types

[38] Both players Hl and Hs always reject an offer of 30 and reject an offer of 50 more than 80 percent of the time.
[39] Recall that the mediator always walks out after messages (h, h).

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 2 does not hold under risk aversion.[40] We did not elicit measures of risk aversion, and cannot rule it out. However, the subjects' other choices in the experiment do not lend it high likelihood. Under DC, demanding 30 as an L is a safe option (as long as the opponent does not walk out, a dominated action). Across subjects, the correlation between the frequency of accepting 50 when Hh in MC and demanding 30 when L in DC is $\widehat{\rho} = -0.06$ (with 95 percent $CI = [-0.296, 0.173]$).[41] Subject-specific risk-aversion seems unlikely to drive the behavior of Hh in MC.

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.

6.3 Neighborhood of best responses

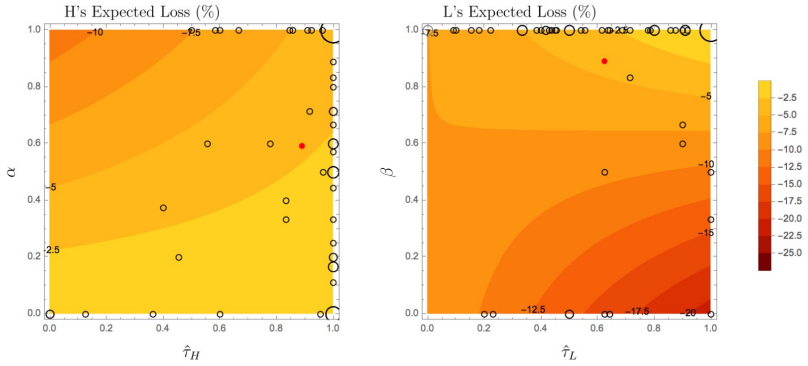
Dominated actions are rare in the data. If we ignore them, each player of given type faces two decisions: the message, $\widehat{\tau}_H$ if H and $\widehat{\tau}_L$ if L, and the acceptance of 50 if H, α, and of 30 if L, β.[42] For each session, we calculated the average strategies played by all the players in the 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 parametrizations (averaged over all sessions). The panels in Figure 8 are contour plots reproducing the loss from not best responding, as a percentage of the maximum possible payoff. The upper panel (a) refers to $q = 1/2$ and the lower panel (b) to $q = 1/3$; in both cases the left panel refers to an

[40] Under the MC 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. The truthful equilibrium where H types always reject 50 ($\alpha_h = 0, \alpha_l = 0, \tau_H = 1, \tau_L = 1$) continues to exist.

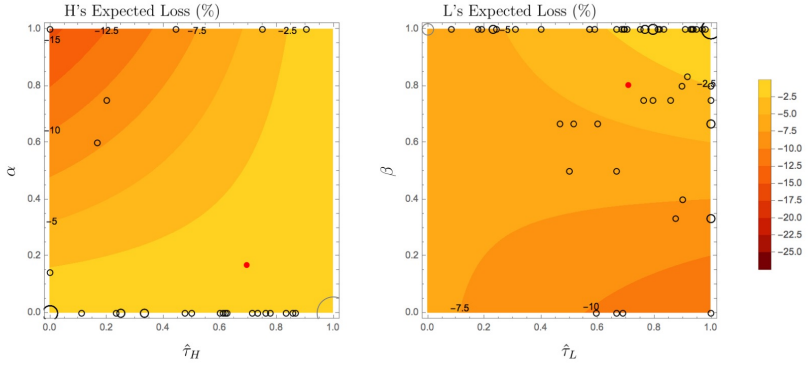
[41] Note that only behavior under $q = 1/2$ is relevant here because subjects make choices under a single parametrization, and we are investigating possible risk aversion for subjects accepting 50 under MC when $q = 1/2$.

[42] Combining α_l and α_h if $q = 1/2$.

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.



(a) $q = 1/2$.



(b) $q = 1/3$.

Figure 8: Losses relative to best responding to the average empirical strategy in the MC treatment. The horizontal axes correspond to messaging strategies; the vertical axes to acceptance strategies; left panels refer to H types; right panels to L types.

In the $q = 1/2$ sessions, there is a 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 for L types, 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. However, the loss from accepting instead is small.

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

The contour plots show that in both parametrizations, 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 individual 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 parametrization and type, random messages and random acceptances; we then ran Kolmogorov-

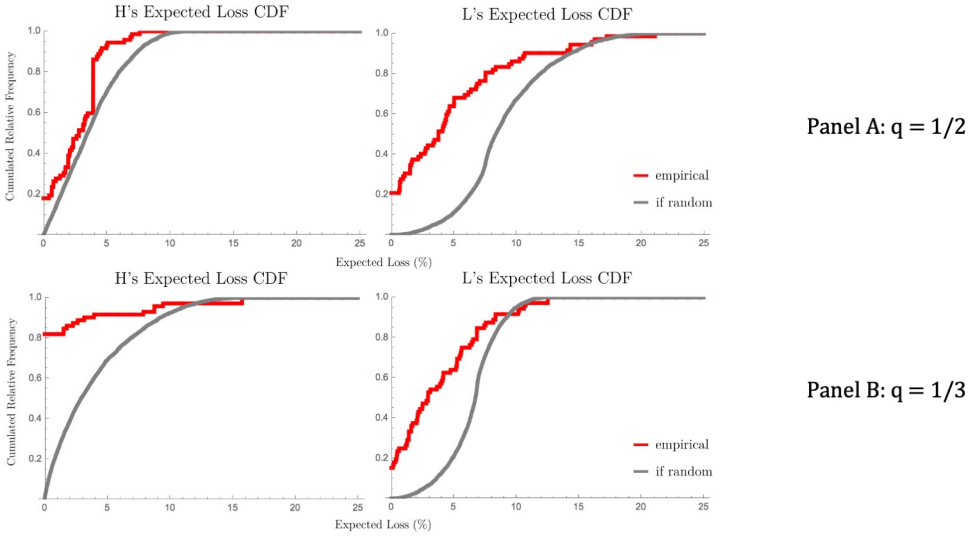


Figure 9: CDF's of losses, given observed play by others.

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.[43] All eight resulting tests strongly reject the hypothesis that subjects' choices were random ($p < 0.001$ in all cases).

Figure 9 compares CDFs of losses, in the data (in red), and under random decision-making (in grey) for each player's type and each of the two parametrizations. 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 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

[43] 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.

responses, are not far from them in payoff space.

7 Adding slackness to incentive constraints

Even if experimental subjects are close to best responding, they are not playing the best equilibrium of the MC program. In the optimal mediation mechanism, incentive constraints are designed to hold weakly: players are brought to the point of indifference, and the ex ante probability of peace is maximized by eliminating any surplus from obeying the mediator. It is natural to ask whether a mechanism with strict costs for deviating from the mediator's recommendations would perform better in the lab. Making obedience more transparently profitable could in principle lead to a higher realized frequency of peace. Blume et al. (2023) follow this reasoning in designing and testing a mechanism for mediated cheap talk that includes slack in the incentive constraints and thus is expressly inefficient.

In this section, we introduce a mechanism with strict constraints and report the results of eight additional experimental sessions run after the original experiment was concluded.

We seek a family of mechanisms that are robust in the following specific sense: in the best equilibrium, equilibrium actions either yield a strictly higher payoff than deviation or are weakly dominant. For both parametrizations, we want to specify a mediation program such that there exists a strict equilibrium where both types are fully truthful ($\tau_H = \tau_L = 1$, $\sigma_H = \sigma_L = 0$), and the mediator's recommendations are obeyed with probability 1 ($\alpha_h = 1$ and $\beta = 1$). Consider the following families of mechanisms:

If $\underline{q = 1/2}$, the mechanisms are indexed by $a \in [0, \frac{3}{8})$ and $b \in [0, \frac{1}{2})$. They are: $r(l,l) = (0.5, 0.5)$; $r(h,l) = \{(0.7, 0.3)$ with probability $5/8+a$, $(0.5, 0.5)$ otherwise$\}$; $r(h,h) = \{(0.5, 0.5)$ with probability $1/2 - b$, w otherwise$\}$.

When $a = b = 0$, the mechanism coincides with the HMS mechanism. A strict equilibrium that satisfies truthfulness and compliance exists for all $\frac{4}{5}a < b < \frac{4}{3}a$. The ex ante probability of peace P equals $\frac{7}{8} - \frac{1}{4}b$.

If $\underline{q = 1/3}$, the mechanisms are indexed by $a \in [0, \frac{3}{4})$ and $b \in [0, 1)$. They are: $r(l,l) = (0.5, 0.5)$; $r(h,l) = \{(0.7, 0.3)$ with probability $3/4 - a$, w otherwise$\}$; $r(h,h) = \{(0.5, 0.5)$ with probability b, w otherwise$\}$.

Again, the mechanism coincides with the HMS mechanism if $a = b = 0$. A strict equilibrium with truthfulness and obedience exists for all $0 < b < \frac{4}{5}a$. The ex ante probability of peace P equals

$(7 - 4a + b)/9$.

We call the modified mediation treatment MCR (Mediated Communication - Robust). We ran four additional experimental sessions for each of the two parametrizations. In both cases, we ran two sessions with order {DC, MCR} and two with order {MCR, DC}. For comparison with the original data, besides setting a and b at positive values, we kept everything else unchanged.[44]

We chose values of a and b that support the desired strict equilibria, lead to simple probabilities of different recommendations (and thus intuitive enough mediation programs), and preserve sizable uncertainty about the opponent's type under $q = 1/2$. With $q = 1/2$, setting $a = 0.175$ and $b = 0.2$, the mediation program becomes:

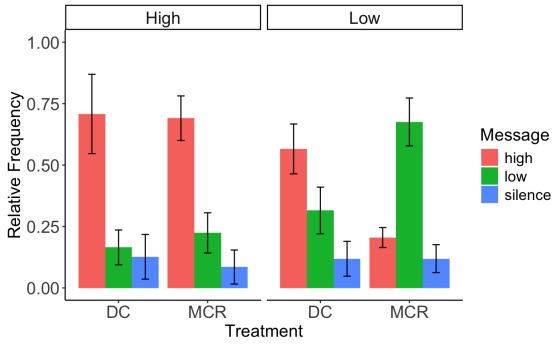
$\underline{q = 1/2}$. $r(l, l) = (0.5, 0.5)$; $r(h, l) = \{(0.7, 0.3)$ with probability 80 percent, $(0.5, 0.5)$ otherwise$\}$; $r(h, h) = \{(0.5, 0.5)$ with probability 30 percent, w otherwise$\}$. The peace probability in the best equilibrium is $P = 0.825$ (compared to $P = 0.875$ in the HMS mechanism).

With $q = 1/3$, setting $a = 0.25$ and $b = 0.125$, the mediation program becomes:

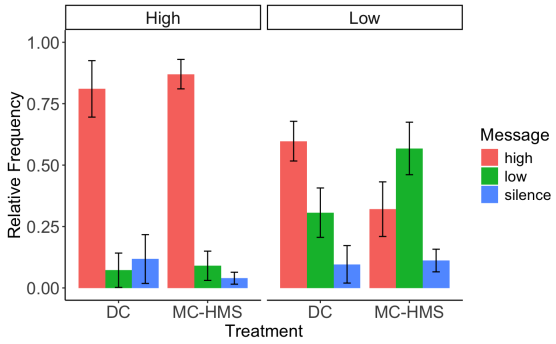
$\underline{q = 1/3}$. $r(l, l) = (0.5, 0.5)$; $r(h, l) = \{(0.7, 0.3)$ with probability 50 percent, w otherwise$\}$; $r(h, h) = \{(0.5, 0.5)$ with probability 12.5 percent (1/8), w otherwise$\}$. The peace probability in the best equilibrium is $P = 0.68$ (compared to $P = 0.778$ in the HMS mechanism).

The amended mechanisms are intriguing, but do not modify the results, relative to the original experiments with the HMS mechanism: as before, mediation increases sincerity for L types, but does not increase peace. Figures 10 and 11 report the results of these auxiliary sessions, compared to the data from the original experiments (here labeled MC-HMS for clarity). Relative to MC-HMS, sincerity under MCR is slightly higher for L types in both parametrizations, and for H types if $q = 1/3$. However, relative to DC, results are unchanged: mediation has a negligible effect on the sincerity of H types and a clear and sizable effect on the sincerity of the L types (Figure 10). At the same time, the impact of mediation on the frequency of peace remains nil at best: we observe a small, insignificant decline relative to DC (Figure 11). Introducing slack in the incentive constraints under mediation does not lead to a decline in conflict, either relative to DC or relative to the previous results under the HMS mechanism.

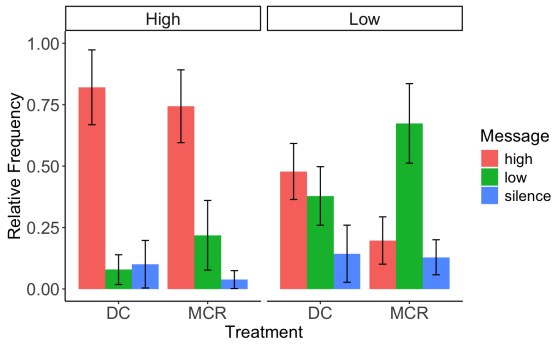
[44] In particular, as in the original sessions, in-between the DC and MCR rounds, we ran a treatment where a subject played the mediator, with no pre-specified program.



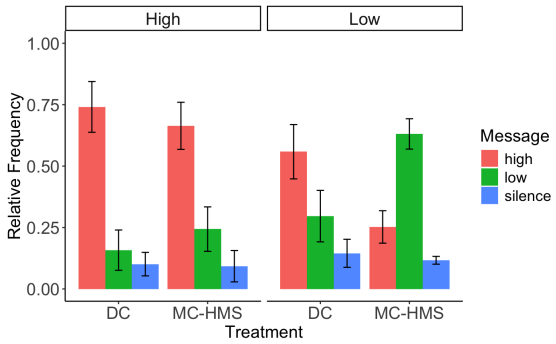
(a) DC and MCR; $q = 1/2$. Auxiliary sessions.



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



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



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

Figure 10: Sincerity: DC and MCR v/s DC and MC-HMS

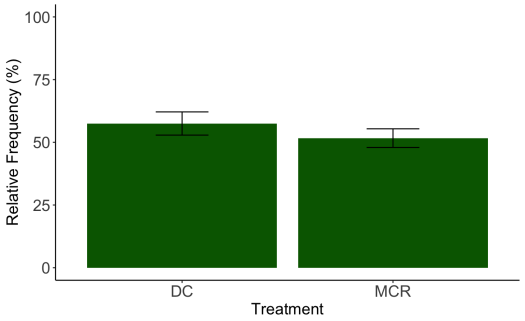
The online appendix (Section B.5.6) presents results on sincerity and peace regressions for the MCR sessions, next to the same regressions for the original sessions with the optimal HMS mechanism. They confirm the conclusions of the previous analysis—like the HMS mechanism, the MCR mechanism helps to obtain more sincerity but not more peace.

We propose that the robust mechanism does not deliver higher peace because, while the best equilibrium is strict, there is still a large multiplicity of equilibria,[45] and the best equilibrium under $q = \frac{1}{2}$ is still fragile to deviations from full sincerity. As Proposition 2R states, for any such deviations, no matter how small, the probability of peace remains bounded away from its value in the best equilibrium:[46]

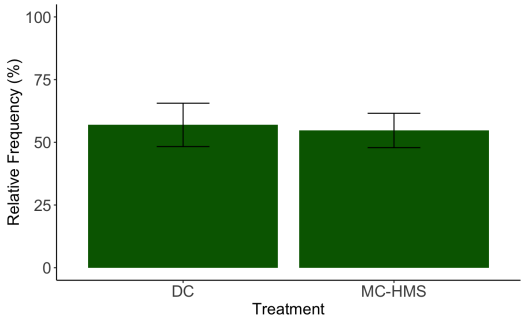
Proposition 2R. *In the robust mechanism for $q = \frac{1}{2}$, for any convergent sequence of equilibria in*

[45] Multiple equilibria continue to exist for both parametrizations, which we report in the online appendix (Section B.3). The equilibria can be visualized at https://www.youtube.com/watch?v=8NIeQ1N_KNo.

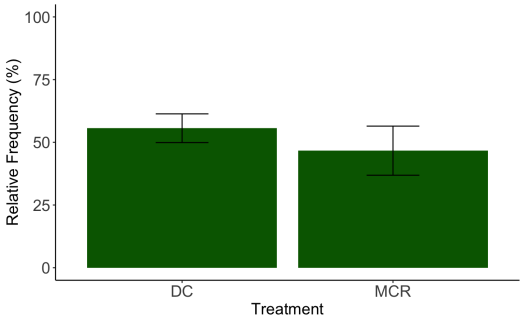
[46] Proposition 2R differs from Proposition 2 because, under the robust mechanism, there exist equilibria in the neighborhood of full sincerity such that $\alpha_h > 0$. However, for all such equilibria α_h is bounded away from 1 and hence the discontinuity in peace is confirmed.



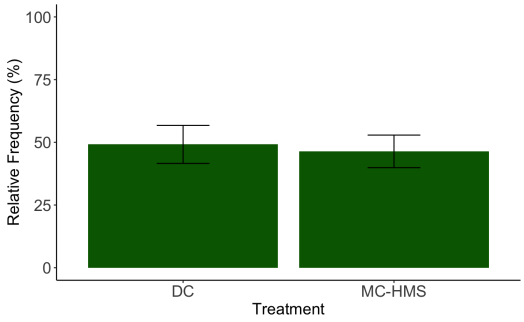
(a) DC and MCR; $q = 1/2$. Auxiliary sessions.



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



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



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

Figure 11: Peace: DC and MCR v/s DC and MC-HMS

which $(\tau_H^t, \tau_L^t) \to (1,1)$ *with either* $\tau_H^t < 1$ *or* $\tau_L^t < 1$ *for all* t, $\lim_{t\to\infty} \alpha_h^t < 1$ *and* $\lim_{t\to\infty} P^t < \frac{7}{8} - \frac{1}{4}b$.

Proof. See online appendix (Section B.4) □

8 Conflict and peace in the absence of communication

As we described in Section 4, we began each experimental session with 10 introductory rounds in which subjects expressed demands without exchanging messages. The design was identical to the DC treatment but without the message stage: subjects expressed their demands; demands were satisfied if compatible; if not, the resource shrank and was shared according to the players' types. We denote these rounds as NCI (no communication - introductory). They provide an intriguing benchmark for conflict and peace in the absence of communication.

We report in this section both the original data and the results of four additional sessions run after the main experiment was concluded. These auxiliary sessions were designed to make the No Communication rounds more comparable to the other treatments: over 20 rounds rather than 10,

and in the middle of the experiment after experience with either DC or MC. We refer to the No Communication treatment from the auxiliary sessions as NC (as opposed to NCI for the introductory rounds).[47]

Recall that in all sets of equilibria we characterized for the DC game, equilibrium messages span both partially informative and fully non-informative messages. Equilibria with fully non-informative messages are equilibria of the game without communication. In addition, because the expected frequency of peace is constant within each equilibrium set, under our selection criteria the theory predicts equal frequencies of peace under DC and under NC: the DC equilibrium frequencies of peace in Table 1 remain equilibrium frequencies of peace under NC. According to the theory, communication per se need not improve outcomes. What did the experimental data show?

In the data collected in the original sessions, the observed frequency of peace in the initial 10 rounds without communication (NCI) 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 $CI = [0.468, 0.587]$), relative to predictions of 0.586 under $q = 1/2$ and 0.444 under $q = 1/3$. Thus peace fits the theory quite well for $q = 1/2$ and is higher than expected under $q = 1/3$.[48]

The four auxiliary sessions that were run later aimed at evaluating the robustness of these findings. All four sessions had $q = 1/2$; two were run in order NCI(10), DC(20), NC(20), MC(20) (with the number of rounds in parenthesis), and two in order NCI(10), MC(20), NC(20), DC(20). As we did for the original data, we denote by NCI the 10 initial training rounds under NC, which we maintained in the auxiliary sessions.[49]

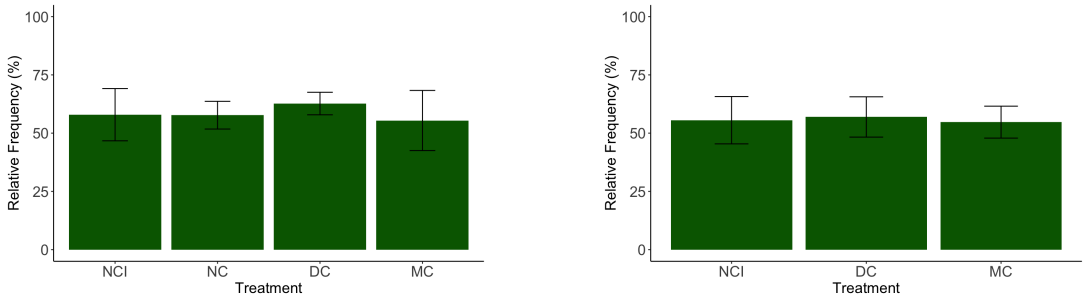
Figure 12 shows the frequency of peace in the auxiliary sessions, on the left, and in the original data, on the right. In both cases, we compare NC to both DC and MC. In the auxiliary sessions, NC does marginally worse in terms of peace, compared to DC, than in the original data, but the effect is small and not significant. The peace frequency is identical between initial NCI rounds and the more numerous NC rounds - the only difference is a slightly narrower confidence interval, as expected. We also find effectively no difference relative to the original data. In our experiment, optimal mediated communication fails to reduce conflict not only relative to direct communication between the two

[47] This section focuses on the frequency of peace. We discuss demand strategies under NC in Section B.5.5 of the online appendix.

[48] The predicted treatment effect, with higher peace under $q = 1/2$, is observed in the data, but with large confidence intervals.

[49] In the treatments involving messaging, Silence was disallowed, but note that this does not affect the NC rounds, and neither does it affect the comparison with the original NCI initial rounds, run before any experience with messaging.

opposite parties but also relative to no communication at all.[50]



(a) NCI, NC, DC and MC; $q = 1/2$. Auxiliary sessions. (b) NCI, DC, and MC; $q = 1/2$. Original sessions.

Figure 12: Peace: NCI, NC, DC, and MC

9 Conclusions

This study analyzes, both theoretically and experimentally, whether a sophisticated mediation algorithm can reduce conflict between two parties who are uncertain about their opponent's strength. Although the mediator has no superior information, no independent resources, and no power to enforce a recommendation, theory predicts that mediation can lead to a strictly lower frequency of conflict than if the two parties communicate directly. We test the optimal mediation protocol in the lab and find that while participants reveal their strength more sincerely to the mediator than to each other, the frequency of conflict is not lower.

In fact, neither mediation nor direct communication between the parties appears to improve outcomes over a scenario in which only demands are exchanged.

Having established this finding, we devote the second part of our analysis to understanding its origins. As typical of mechanism design, the optimal mediation program has multiple equilibria and the multiplicity is partly responsible for the result. Designing mechanisms with a restricted set of equilibria or compelling criteria for equilibrium selection remains a clear priority, especially in applied research.

However, two other factors specific to optimal mediation also play a role. First, theory suggests that the superiority of mediation relative to direct communication between the parties is tied to

[50] In the auxiliary sessions, DC does slightly better also relative to MC, but again the difference is small and statistically insignificant. The observation also confirms that the comparison between DC and MC is not affected substantively by the treatment run in-between the two (NC in the auxiliary sessions or mediation by a subject without a pre-specified protocol in the original sessions).

the mediator's ability to leave the parties unsure of their opponent's strength, that is, to *obfuscate* the message received by the other party. We find that it is exactly when obfuscation is part of the optimal mechanism that the 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 downward jump in the frequency of peace. The jump occurs because, when there is obfuscation, the positive probability of lies makes non-compliance with some of the mediator's recommendations profitable. In the lab, sincerity is not perfect and, as predicted, neither is compliance.

It is important to note that neither the multiple equilibria problem nor the fragility of obfuscation is a result of the weak incentives operating under the optimal mechanism. Theory tells us that both are preserved under mechanisms with strict incentive constraints. We ran additional experimental sessions with such a mechanism. The frequency of conflict is indistinguishable from what we find under the optimal mechanism (and thus from what we find under direct communication).

Second, in the lab we see deviations from equilibrium. Given the difficult game, noise in the subjects' behavior is not surprising. Rather, what is interesting is that behavior is far from erratic. We find that individual deviations from best responding to the empirical strategies others play in the lab cause each subject only minor individual losses—typically less than 5 percent of their payoff. The problem is that such deviations have significant negative repercussions on the overall frequency of conflict and thus on the payoff of others.

Looking at empirical best responses, one additional finding confirms the fragility of optimal mediation with obfuscation: empirical best responses correspond to sincerity and compliance in the absence of obfuscation, but do not when the mediation mechanism includes obfuscation. The incentive to reject the mediator's recommendation identified by the theory under fully rational behavior in the presence of lies is reflected in the empirical incentives faced by participants in the lab.

We come away from the experiment with several open directions for future research. First, as mentioned, when studying mediation mechanisms it remains important to make progress on equilibrium uniqueness or selection.

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. Going beyond the mechanisms with slack we tested in Section 7, it would be good to design mediation protocols that are strategically simple (as discussed in Li, 2017 or Börgers

and Li, 2019, for example) or that rely on boundedly rational thinking (as in Kneeland, 2022 or de Clippel et al., 2019, for example).

These directions call both for theoretical progress and for lab experiments—experiments disciplined by a very precise theory and a high level of control. If the final goal, however, is to identify potential tools for interventions, eventually we will need to go beyond the lab to the field. Besides the loss of control, field data will reintroduce the psychological aspects that our lab studies can abstract from. A careful analysis of data from existing Alternative Dispute Resolution procedures would be a desirable first step.

We conclude by underlining again that in this study we have implemented the optimal mediation mechanism as a computer-run algorithm. Such a choice allows us to control the design of the mechanism and to test the theory cleanly. From an applied perspective, it is in line with the industry's increasing reliance on algorithmic mediation programs, and we hope that the weaknesses we identify may be instructive for practical applications. However, it is also important to study mediation when it is provided by a human mediator, or, in an experiment, by one of the experimental participants. In the most natural design, there is no commitment, complicating substantially the theoretical analysis. At the same time, with the behavior of the mediator more difficult to predict, the experimental game becomes very difficult for the subjects to process. We report initial results in Casella, Friedman, and Perez Archila (2020), but more work is needed to overcome these obstacles productively.

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.

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

Blume, A., E. K. Lai and W. Lim, 2019, "Eliciting private information with noise: The case of randomized response", *Games and Economic Behavior*, 113, 356-380.

Blume, A., E. K. Lai and W. Lim, 2023, "Mediated Talk: An Experiment", *Journal of Economic Theory*, 208.

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.

Casella, A., E. Friedman and M. Perez Archila, 2020, "Mediating Conflict in the Lab", NBER W.P. No. 28137, Cambridge, Ma.

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., 2021, "Mediation in Reputational Bargaining", *American Economic Review*, 111, 2444-72.

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, 2022, "Rules and Commitment in Communication", *Econometrica*, 90, 2283-2318.

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 Negoti-

ation", *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.

John, L. K., G. Loewenstein G., A. Acquisti, and J. Vosgerau, 2018, "When and why randomized response techniques (fail to) elicit the truth", *Organizational Behavior and Human Decision Processes* 148, 101–123.

Kneeland, T., 2022, "Mechanism Design with Level-k Types: Theory and an Application to Bilateral Trade", *Journal of Economic Theory*, 201.

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

Krishna, V, and J. Morgan, 2004, "The art of conversation: eliciting information from experts through multi-stage communication", *Journal of Economic Theory*, 117, 147–179.

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.

Ljungqvist, L., 1993, "A Unified Approach to Measures of Privacy in Randomized Response Models: A Utilitarian Perspective", *Journal of the American Statistical Association*, 88, 97-103.

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., 1978, "Refinements of the Nash Equilibrium Concept", *International Journal of Game Theory*, 15, 133-154.

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.

Warner, S., 1965, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias", *Journal of the American Statistical Association*,60, 63-69.

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.

A Appendix

A.1 Direct 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 between 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:*

$$\delta_\theta(H, m, m') = 1 \text{ for all } m, m'$$
$$\delta_\theta(L, m, m') = 1 - \delta_{1-\theta}(L, m, m') = 2\left(1 - \frac{1-\theta}{\theta(1-\pi_m)}\right)$$

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

$$\tau_H \geq \max\left\{\left(\frac{3\theta-2}{2(1-\theta)}\right)\left(\frac{1-q}{q}\right)(1-\sigma_L-\tau_L), 1-\sigma_H-\left(\frac{2\theta-1}{1-\theta}\right)\left(\frac{1-q}{q}\right)\tau_L\right\} \quad (3)$$
$$\tau_H \leq \min\left\{\left(\frac{2\theta-1}{1-\theta}\right)\left(\frac{1-q}{q}\right)(1-\sigma_L-\tau_L), 1-\sigma_H-\left(\frac{3\theta-2}{2(1-\theta)}\right)\left(\frac{1-q}{q}\right)\tau_L\right\}$$
$$\sigma_H \in \left[\left(\frac{3\theta-2}{2(1-\theta)}\right)\left(\frac{1-q}{q}\right)\sigma_L, \left(\frac{2\theta-1}{1-\theta}\right)\left(\frac{1-q}{q}\right)\sigma_L\right]$$

$(\tau_H + \sigma_H) \in (0,1)$.

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:*

$$\delta_\theta(H, m, m') = 1 \text{ for all } m, m'$$
$$\delta_{1/2}(L, m, m') = 1 \text{ for all } m, m'.$$

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

$$\begin{aligned} \tau_H &\in \left[1 - \sigma_H - \left(\frac{2\theta-1}{1-\theta}\right)\left(\frac{1-q}{q}\right)\tau_L, \left(\frac{2\theta-1}{1-\theta}\right)\left(\frac{1-q}{q}\right)(1-\sigma_L-\tau_L)\right] \\ \sigma_H &\leq \left(\frac{2\theta-1}{1-\theta}\right)\left(\frac{1-q}{q}\right)\sigma_L \qquad (4) \\ (\tau_H + \sigma_H) &\in (0,1). \end{aligned}$$

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

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

B, C, R and E: (1) Both E and R demand θ, B demands $1-\theta$ and C demands θ if $\pi_l \leq (2\theta-1)/\theta$ and $\pi_h \geq (3\theta-2)/\theta$. (2) Both E and R demand θ, C demands $1-\theta$ and B demands θ if $\pi_h \leq (2\theta-1)/\theta$ and $\pi_l \geq (3\theta-2)/\theta$. (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 $\pi_l \leq 2\theta-1$ and $\pi_h \leq 2\theta-1$.

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:

$$\begin{aligned}
S_L(l) &= [(1-q)\tau_L + q(1-\tau_H)]\widehat{S}_A + [q\tau_H + (1-q)(1-\tau_L)]\widehat{S}_B \\
S_L(h) &= [(1-q)\tau_L + q(1-\tau_H)]\widehat{S}_C + [q\tau_H + (1-q)(1-\tau_L)]\widehat{S}_D \\
S_H(h) &= [(1-q)\tau_L + q(1-\tau_H)]\widehat{S}_E + [q\tau_H + (1-q)(1-\tau_L)]\widehat{S}_F \\
S_H(l) &= [(1-q)\tau_L + q(1-\tau_H)]\widehat{S}_G + [q\tau_H + (1-q)(1-\tau_L)]\widehat{S}_R.
\end{aligned} \tag{5}$$

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 (5) 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:

$$\begin{aligned} \pi_h &\in [(3\theta-2)/\theta, (2\theta-1)/\theta] \\ \pi_l &\in [(3\theta-2)/\theta, (2\theta-1)/\theta]. \end{aligned} \tag{6}$$

For any $\tau_L \in (0,1)$, ruling out silence, conditions (6) correspond to the restrictions on τ_H identified in Proposition A1 (inequalities (3), 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,[51] 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:

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

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:

$$\sigma_H \in \left[\frac{3\theta - 2}{2(1-\theta)}\left(\frac{1-q}{q}\right)\sigma_L, \frac{2\theta - 1}{1-\theta}\left(\frac{1-q}{q}\right)\sigma_L\right] \tag{7}$$

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

Condition (7) must be satisfied, together with conditions (6). 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:

[51] 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.

$$S_L(l) = [(1-q)\tau_L + q(1-\sigma_H-\tau_H)]\widehat{S}_A + [(1-q)\sigma_L + q\sigma_H]\widehat{S}_{A_{s2}} + [q\tau_H + (1-q)(1-\sigma_L-\tau_L)]\widehat{S}_B$$
$$S_L(h) = [(1-q)\tau_L + q(1-\sigma_H-\tau_H)]\widehat{S}_C + [(1-q)\sigma_L + q\sigma_H]\widehat{S}_{D_{s2}} + [q\tau_H + (1-q)(1-\sigma_L-\tau_L)]\widehat{S}_D \quad (8)$$
$$S_L(s) = [(1-q)\tau_L + q(1-\sigma_H-\tau_H)]\widehat{S}_{A_{s1}} + [(1-q)\sigma_L + q\sigma_H]\widehat{S}_{L_{ss}} + [q\tau_H + (1-q)(1-\sigma_L-\tau_L)]\widehat{S}_{Ds1}$$

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 (4) 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 (3) 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.

A.2 Multiple equilibria under MC (at the experimental parameter values)

We characterize players' equilibrium strategies keeping fixed the mediator's mechanism as programmed under MC. 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 7 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 B.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.

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

$q = 1/2$		$q = 1/3$	
(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)	$\alpha_l = 0, \beta = 1, \hat{\tau}_L \in (0,1)$, $\hat{\tau}_H = 1/3 + (2/3)\hat{\tau}_L$
(iii)	$\alpha_l = 0, \alpha_h = 0, \hat{\tau}_L = 0, \hat{\tau}_H \in [1/6, 4/15]$	(iii)	$\alpha_l = 0, \beta = 4/7, \hat{\tau}_L \in (0,1)$, $\hat{\tau}_H = 1/3 - \hat{\tau}_L/3$
(iv)	$\alpha_l = 0, \alpha_h = 0, \beta = 1, \hat{\tau}_L \in (0,1)$, $\hat{\tau}_H = 4/15 + (6/15)\hat{\tau}_L$	(iv)	$\alpha_l = 0, \hat{\tau}_L = 0, \hat{\tau}_H \leq 1/3$
(v)	$\alpha_l = 0, \alpha_h = 0, \beta = 1, \hat{\tau}_L = 1, \hat{\tau}_H \in [2/3, 1)$		
(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 7: MC: Equilibria in undominated strategies

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:[52]

$$\Pr(j \text{ is } L | h_j) = \frac{2(1 - \tau_L)}{2(1 - \tau_L) + \tau_H}. \tag{9}$$

In any equilibrium in which all accept 70, Ll player i will accept 30 with positive probability only if $EU_{Ll}(accept\ 30) \geq EU_{Ll}(reject\ 30)$ where $EU_{Ll}(accept\ 30) = 30$ and $EU_{Ll}(reject\ 30) = 35 \Pr(j = L | h_j)$. Substituting (9):

$$3\tau_H = 1 - \tau_L \Longrightarrow \beta \in [0,1] \text{ and } 3\tau_H < 1 - \tau_L \Longrightarrow \beta = 0,\ 3\tau_H > 1 - \tau_L \Longrightarrow \beta = 1. \tag{10}$$

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.

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

[52]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 (9) is well-defined. A similar observation applies to other posterior probabilities below, and is not repeated.

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

$$\Pr(j\ accepts\ and\ is\ H|50, l_j) = \frac{\alpha_l(1-\tau_H)}{(1-\tau_H)+2\tau_L}, \tag{11}$$
$$\Pr(j\ accepts\ and\ is\ L|50, l_j) = \Pr(j\ is\ L|\ l_j) = \frac{2\tau_L}{2\tau_L+(1-\tau_H)}.$$

Hence:

$$15\alpha_l(1-\tau_H) = 20(2\tau_L) \Longrightarrow \alpha_l \in [0,1] \text{ and} \tag{12}$$
$$15\alpha_l(1-\tau_H) > 20(2\tau_L) \Longrightarrow \alpha_l = 1,\ 15\alpha_l(1-\tau_H) < 20(2\tau_L) \Longrightarrow \alpha_l = 0.$$

As above, since we are considering the acceptance decision for 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:

$$EU_H(h) = (1/3)\left[\tau_H 35 + (1-\tau_H)(35/4 + 35(3/4))\right] + (2/3)\left[\tau_L 70 + (1-\tau_L)70\right] = (1/3)35 + (2/3)70 \tag{13}$$

$$EU_H(l) = (1/3)\left[\tau_H 35 + (1-\tau_H)(\alpha_l^2 50 + (1-\alpha_l^2)35)\right] + (2/3)\left[\tau_L(\alpha_l 50 + (1-\alpha_l)70) + (1-\tau_L)70\right]$$
$$EU_L(l) = (1/3)[\tau_H(3/4)\beta 30 + (1-\tau_H)\alpha_l 50] + (2/3)[\tau_L 50 + (1-\tau_L)(35/4 + (3/4)(30\beta + 35(1-\beta)))]$$
$$EU_L(h) = (1/3)[\tau_H(0) + (1-\tau_H)(0)] + (2/3)[\tau_L(35/4 + (3/4)(\beta 70 + (1-\beta)35)) + (1-\tau_L)35].$$

Four conditions, ((9), (11), 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 (12) and (10). 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 7. 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 MC, 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: $\Pr(j \text{ is } L|\widehat{h}_j) = \frac{\Pr(\widehat{h}_j|j \text{ is } L)\Pr(L)}{\Pr(\widehat{h}_j|j \text{ is } L)\Pr(L)+\Pr(\widehat{h}_j|j \text{ is } H)\Pr(H)} = \frac{[1-\tau_L-\sigma_L+(1/3)\sigma_L](2/3)}{[1-\tau_L-\sigma_L+(1/3)\sigma_L](2/3)+[\tau_H+(1/3)\sigma_H](1/3)} = \frac{2(1-\widehat{\tau}_L)}{2(1-\widehat{\tau}_L)+\widehat{\tau}_H}$. With a change in variable, the formula is identical to (9). 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.

For online publication

B Online appendix

B.1 Multiple equilibria under MC: 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 $q = 1/2$:

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

Ll accepts with positive probability if:

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

or:

$$6\tau_H > 1-\tau_L \Longrightarrow \beta = 1,\ 6\tau_H < 1-\tau_L \Longrightarrow \beta = 0, \text{ and } 6\tau_H = 1-\tau_L \Longrightarrow \beta \in [0,1]. \tag{14}$$

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

The relevant posterior probability is:

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

where:

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

and:

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

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

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

and

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

Hh will accept 50 with positive probability if:

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

or:

$$\begin{aligned}15(4\tau_H\alpha_h + 3(1-\tau_H)\alpha_l) = 20(4(1-\tau_L) + 3\tau_L) \Longrightarrow \alpha_h \in [0,1] \\ 15(4\tau_H\alpha_h + 3(1-\tau_H)\alpha_l) < 20(4(1-\tau_L) + 3\tau_L) \Longrightarrow \alpha_h = 0, \\ 15(4\tau_H\alpha_h + 3(1-\tau_H)\alpha_l) > 20(4(1-\tau_L) + 3\tau_L) \Longrightarrow \alpha_h = 1.\end{aligned} \tag{16}$$

Condition (15) 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:

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

and will accept 50 with positive probability if:

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

or:

$$\begin{aligned} 15(3\tau_H\alpha_h + 8(1-\tau_H)\alpha_l) = 20(3(1-\tau_L) + 8\tau_L) &\implies \alpha_l \in [0,1] \\ 15(3\tau_H\alpha_h + 8(1-\tau_H)\alpha_l) < 20(3(1-\tau_L) + 8\tau_L) &\implies \alpha_l = 0 \\ 15(3\tau_H\alpha_h + 8(1-\tau_H)\alpha_l) > 20(3(1-\tau_L) + 8\tau_L) &\implies \alpha_l = 1. \end{aligned} \tag{17}$$

Conditions (14), (16), and (17) 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 (14). 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:

$$\begin{aligned} EU_H(h) &= (1/2)[\tau_H 35 + (1-\tau_H)((5/8)35 + (3/8)(50\alpha_l + 35(1-\alpha_l)))] + (1/2)70 \\ EU_H(l) &= (1/2)[\tau_H 35 + (1-\tau_H)(50\alpha_l^2 + 35(1-\alpha_l^2))] + \\ &\quad (1/2)[\tau_L(\alpha_l 50 + (1-\alpha_l)70) + (1-\tau_L)((5/8)70 + (3/8)(50\alpha_l + 70(1-\alpha_l)))] \\ EU_L(l) &= (1/2)[\tau_H(5/8)30\beta + (1-\tau_H)50\alpha_l] + (1/2)[\tau_L 50 + (1-\tau_L)((5/8)(30\beta + 35(1-\beta)) + (3/8)50)] \\ EU_L(h) &= (1/2)[(1-\tau_H)((3/8)50\alpha_l] + \\ &\quad (1/2)[\tau_L((5/8)(70\beta + 35(1-\beta)) + (3/8)50) + (1-\tau_L)(50/2 + 35/2)]. \end{aligned} \tag{18}$$

As before, four conditions, (17), (14), 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 B1. *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 (17):

$$6(1-\tau_H)\alpha_l = 3 + 5\tau_L \Longrightarrow \tau_H = 1 - \left(\frac{3+5\tau_L}{6\alpha_l}\right). \tag{19}$$

Substituting (19) in (18), 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 (19) is violated. Thus $\alpha_l \in (0,1)$ is impossible.[53]

(2) Suppose then $\alpha_l = 1$. From (17), it follows that:

$$\tau_H \leq (1/2) - (5/6)\tau_L. \tag{20}$$

Note that there cannot be an equilibrium with $\alpha_l = 1$ if L prefers sincerity and thus $\tau_L = 1$. From (18):

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

an expression that is minimal when τ_H is maximal. By (20), 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 (20) there cannot be an equilibrium with $\alpha_l = 1$, and the Lemma is proven. □

Proposition 2 and Lemma B1 establish $\alpha_l = 0$ and, unless $\tau_L = 1$ and $\tau_H = 1$, $\alpha_h = 0$. Studying

[53] 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 (17) an H player who lied would still reject 50.

(18) and (14), we can identify the full set of equilibria:[54]

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

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

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

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

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

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

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

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.[55] 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

[54] 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.

[55] 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 (14) 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.

B.2 Trembling-hand perfection

If $(2\theta - 1) < q < (2\theta - 1)/\theta$, the HMS equilibrium is (extensive-form) 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:

$$(1/2 - \theta/2)q\left[q_H\tau_H\alpha_h^{1/2} + q_M(1-\tau_H)\alpha_l^{1/2}\right] \geq (\theta - 1/2)(1-q)\left[q_H(1-\tau_L)\beta_h^{1/2} + q_M\tau_L\beta_l^{1/2}\right] \quad (21)$$

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:

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

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

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

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

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

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 constants $\{t_H, t_L, a_h^{1/2}, a_h^{\theta}, a_l^{1/2}, a_l^{1-\theta}, b_l^{1/2}, b_l^{1-\theta}, b_h^{1/2}, b_h^{\theta}\}$ such that:

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

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 parametrization:

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 (21) 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 (21) is then $q_M q_H[\alpha_h^{1/2}[\tau_H + (1-\tau_H)] > q_M q_H[\beta_h^{1/2}(1-\tau_L) + \beta_l^{1/2}\tau_L)]$ 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. Note that, because this is a statement about trembles across distinct information sets, it does imply a violation of (extensive-form) proper equilibrium (Myerson 1978), which imposes restrictions for a given information set.

B.3 Adding slack to the incentive constraints under MC

For $q = \frac{1}{2}$, we consider the following family of mechanisms, indexed by $a \in [0, \frac{3}{8})$ and $b \in [0, \frac{1}{2})$.

- (l, l) offered (50,50) w.p. 1
- (h, l) offered (70,30) w.p. $\frac{5}{8} + a$
- (h, l) offered (50,50) w.p. $\frac{3}{8} - a$
- (h, h) offered (50,50) w.p. $\frac{1}{2} - b$
- (h, h) offered w w.p. $\frac{1}{2} + b$

When $a = b = 0$, the mechanism coincides with that in HMS. Hence we are generalizing the mechanism in HMS. Suppose $\frac{4}{5}a < b < \frac{4}{3}a$.

(1) The following are equilibria:

(i) $\alpha_h = 1$, $\beta = 1$, $\tau_L = 1$, $\tau_H = 1$; $P = \frac{7}{8} - \frac{1}{4}b$

(ii) $\alpha_h = 0$, $\beta = 1$, $\tau_L = 1$, $\tau_H = 1$; $P = \frac{1}{4} + \frac{1}{2}(5/8 + a)$

(iii) $\alpha_l = 0$, $\alpha_h = 0$, $\beta = 1$, $\tau_L = 1$, $\tau_H \in [\frac{6.25+10a}{9.375+15a}, 1]$; $P = \frac{1}{4} + \frac{1}{2}(5/8 + a)\tau_H$

(iv) $\alpha_l = 0$, $\alpha_h = 0$, $\beta = 1$, $\tau_L \in (0, 1)$, $\tau_H = \frac{3.75\tau_L + 7.5b\tau_L + 2.5 - 7.5b + 10a}{9.375+15a}$;

$P = \frac{1}{4}(1-(1-\tau_L)^2(1/2+b)) + \frac{1}{2}\tau_H\tau_L(5/8+a)$

(v) $\alpha_l = 0$, $\alpha_h = 0$, $\tau_L = 0$, $\tau_H \in [0, \frac{2.5-7.5b+10a}{9.375+15a}]$; $P = \frac{1}{4}(\frac{1}{2}-b)$.

(2) Equilibrium (i) is strict: truthful H strictly prefers accepting 50, L strictly prefers telling the truth, and H strictly prefers telling the truth and then complying rather than lying with positive probability and rejecting the recommendation.

Part (1) characterizes equilibria. Equilibrium (i) is the generalization of the optimal equilibrium in HMS, but is strict when $\frac{4}{5}a < b < \frac{4}{3}a$. The full set of equilibria (i)-(v) generalize those characterized in the paper (e.g., represented in the 3D graph of Figure 6. Part (2) gives the three incentive constraints that now hold strictly as part of equilibrium (i), one for L and two for H (including ruling out double deviation)).

We can make precise how the loss from deviation depends on a and b. Call R_L the percentage loss from deviating from full truthfulness for a player of type L, R_H the percentage loss for an H player from refusing 50, and R_{H2} H's percentage loss from double deviation (sending message l and then refusing the mediator's recommendation). Then:

$$R_L = \frac{(20b-16a)}{(35-8a)}$$
$$R_H = \frac{(28-56b)}{(7-8b-a)} - 4$$
$$R_{H2} = \frac{(4a-3b)}{(21+4a-3b)}.$$

For $q = \frac{1}{3}$, we consider the following family of mechanisms, indexed by $a \in [0, \frac{3}{4})$ and $b \in [0, 1)$.

- (l, l) offered (50,50) w.p. 1
- (h, l) offered (70,30) w.p. $\frac{3}{4} - a$
- (h, l) offered w w.p. $\frac{1}{4} + a$
- (h, h) offered (50,50) w.p. b
- (h, h) offered w w.p. $1 - b$

When $a = b = 0$, the mechanism again coincides with that in HMS. Suppose $0 < b < \frac{4}{5}a$.

(1) The following are equilibria:

(i) $\alpha_h = 1$, $\beta = 1$, $\tau_L = 1$, $\tau_H = 1$; $P = \frac{7}{9} - \frac{4}{9}a + \frac{1}{9}b$

(ii) $\alpha_h = 0$, $\beta = 1$, $\tau_L = 1$, $\tau_H = 1$; $P = \frac{7}{9} - \frac{4}{9}a$

(iii) $\alpha_h = 0$,$\alpha_l = 0$, $\beta = 1$, $\tau_L \in (0,1)$, $\tau_H = \frac{\tau_L(5-20a-10b)+2.5+10b-\frac{10}{3}a}{7.5-10a}$;

$P = \frac{4}{9}(\tau_L^2 + 2\tau_L(1-\tau_L)(3/4-a)) + \frac{4}{9}\tau_H\tau_L(3/4-a)$.

(2) Equilibrium *(i)* is strict: both H and L strictly prefer telling the truth and, conditional on having been truthful, prefer complying with the recommendation; H prefers truth and compliance to lying and rejecting the recommendation.

For $q = 1/3$ as well, we can specify percentage losses from deviation from equilibrium, as functions of parameters a and b. We have:

$$R_L = \frac{(16a-20b)}{(49-12a)}$$
$$R_H = \frac{3}{10}$$
$$R_{H2} = \frac{(3b)}{(35+3b)}.$$

B.4 Proof of Proposition 2R

Proof. Let $(\tau_H^t, \tau_L^t) \to (1,1)$ with either $\tau_H^t < 1$ or $\tau_L^t < 1$ for all t. It must be that, for sufficiently high t, $\beta^t = 1$ because an L messaging l offered 30 believes the opponent is H with probability close to 1. Similarly, for sufficiently high t, if $\tau_H^t < 1$, it must be that $\alpha_l^t = 0$ because an H messaging l offered 50 believes that the opponent is L with probability close to 1.[56] Suppose, for contradiction, that $\alpha_h^t \to 1$. In this case, equilibrium t would approach the optimal sincere equilibrium as $t \to \infty$. But then, by continuity, because the optimal sincere equilibrium is strict, for sufficiently high t, it cannot be that either of τ_H^t or τ_L^t are strictly less than 1—a contradiction. Hence $\lim_{t\to\infty} \alpha_h^t =: \bar{\alpha}_h < 1$. The limiting level of peace is $(1-q)^2 + 2q(1-q)(1-(\frac{3}{8}-a)(1-\bar{\alpha}_h)) + q^2(\frac{1}{2}-b)\bar{\alpha}_h^2 < \frac{7}{8} - \frac{1}{4}b$, where the inequality follows from substituting $q = \frac{1}{2}$ and the fact that the left-hand side is strictly maximized at $\bar{\alpha}_h = 1$. □

[56] If $\tau_H^t = 1$, then any α_l^t can be supported for some off-path belief, but this will have no effect on any incentive constraints or the probability of peace.

B.5 Additional experimental results

B.5.1 Sincerity and information transmission: Kullback-Leibler measures

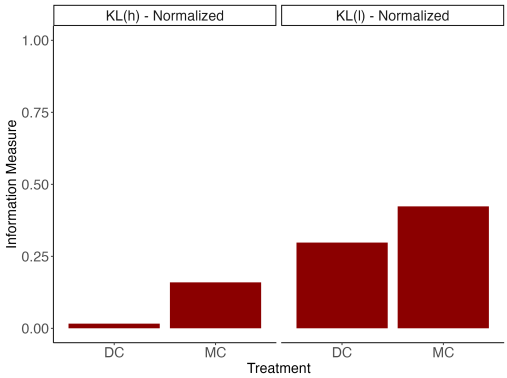
As noted in the text, how much information a message transmits depends on the use of that same message by the opposite type. We can use the Kullback-Leibler (KL) measure of dispersion to generate a summary indicator of the impact of a message on the posterior probability of a given type, relative to 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:

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

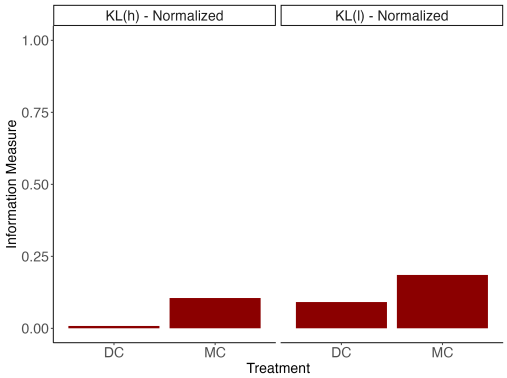
$$KL(l) = \Pr(H|l) \log\left(\frac{\Pr(H|l)}{\Pr(H)}\right) + \Pr(L|l) \log\left(\frac{\Pr(L|l)}{\Pr(L)}\right).$$

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 13 reports, for the two treatments and two parametrizations, the corresponding KL measures for messages l and h, expressed as fractions of the maximum value for each parametrization and averaged over the relevant sessions.



(a) q=1/2



(b) q=1/3

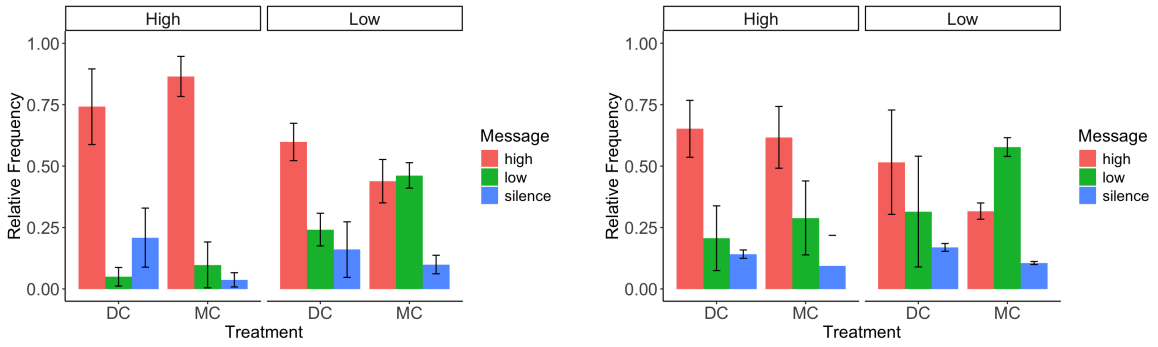
Figure 13: KL measures

It remains true that the treatment conveying most information is MC, although the lesson from the KL measures is more nuanced than Figure 2 in the text and Figure 14 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 parametrizations. 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 $q = 1/2$ parametrization.

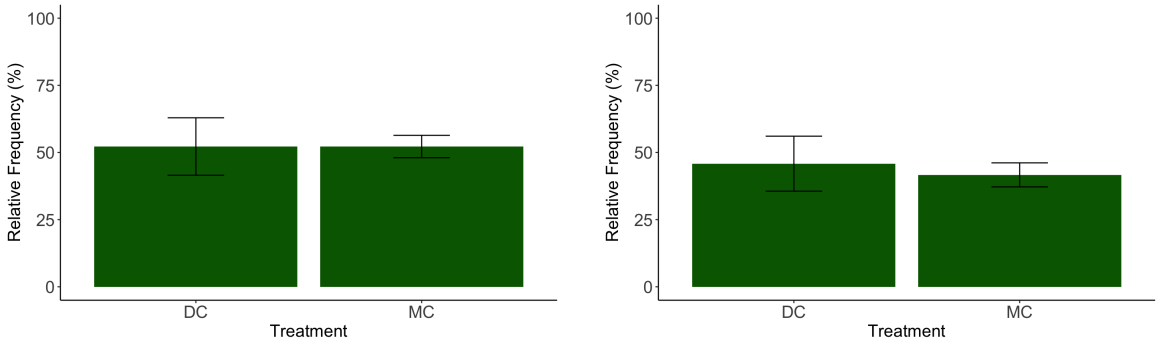
B.5.2 Sincerity and peace in DC and MC, between subjects

We report in Figures 14 and 15 evidence on sincerity and the frequency of peace in DC and MC, comparing sessions where the two treatments are played by subjects in rounds 11-30, i.e., just after the NCI introductory rounds, and with the same experience.



(a) $q = 1/2$ (b) $q = 1/3$

Figure 14: DC and MC: sincerity with equal experience.



(a) $q = 1/2$ (b) $q = 1/3$

Figure 15: DC and MC: frequency of peace with equal experience.

B.5.3 Peace regressions with full interactions terms and by treatments

We report in Table 8 the results of a linear regression of the frequency of peace on treatment, types-pair, order, and round as in Table 5 in the text but with the full set of interaction terms (with the regression specialized by parametrization). The substantive results are unchanged.[57]

In the text, we concluded the description of Table 5 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 9 below. Recall that although higher frequency of peace when the probability of H type realizations is higher appears counterintuitive, the result is in line with the theory. It is predicted in the HMS equilibrium under MC and in all the equilibria we characterize for DC.

[57] We find a positive effect on peace from running the treatment second, but only for $L - L$ pairs and only for $q = 1/3$.

	Dependent variable:	
	Peace	
	$q = 1/2$	$q = 1/3$
MC treatment	−0.027	−0.014
	(0.063)	(0.010)
Pair type *H-L*	0.282***	0.321***
	(0.108)	(0.050)
Pair type *L-L*	0.631***	0.593***
	(0.091)	(0.065)
Second treatment	0.029	−0.008
	(0.063)	(0.010)
Round	−0.005	−0.001
	(0.005)	(0.001)
MC treatment × Pair type *H-L*	0.008	−0.004
	(0.055)	(0.034)
MC treatment × Pair type *L-L*	0.025	−0.005
	(0.076)	(0.041)
Pair type *H-L* × Second treatment	0.043	0.020
	(0.052)	(0.034)
Pair type *L-L* × Second treatment	0.029	0.137***
	(0.074)	(0.044)
Pair type *H-L* × Round	0.005	−0.001
	(0.006)	(0.005)
Pair type *L-L* × Round	0.003	0.006
	(0.004)	(0.005)
Constant	0.265***	0.041**
	(0.093)	(0.021)
Observations	1,440	1,440
Note:	*p<0.1; **p<0.05; ***p<0.01	

The excluded (default) category is the DC treatment, and the first of the two treatments in the session (DC or MC). When looking at different pair types, the default pair is *H-H*. Round refers to the round number within the treatment. Standard errors are clustered at the session level.

Table 8: Peace with interactions.

	Dependent variable:	
	Peace	
	DC	MC
Parameter q=1/2	0.200***	0.202***
	(0.043)	(0.034)
Second treatment	0.068	0.052
	(0.047)	(0.038)
Pair type *H-L*	0.336***	0.344***
	(0.032)	(0.030)
Pair type *L-L*	0.707***	0.721***
	(0.044)	(0.037)
Round	−0.002	0.001
	(0.002)	(0.002)
Constant	0.015	−0.040
	(0.052)	(0.038)
Observations	1,440	1,440
R^2	0.248	0.271
Adjusted R^2	0.245	0.268
Residual Std. Error	0.434 (df = 1434)	0.428 (df = 1434)
Note:		*p<0.1; **p<0.05; ***p<0.01

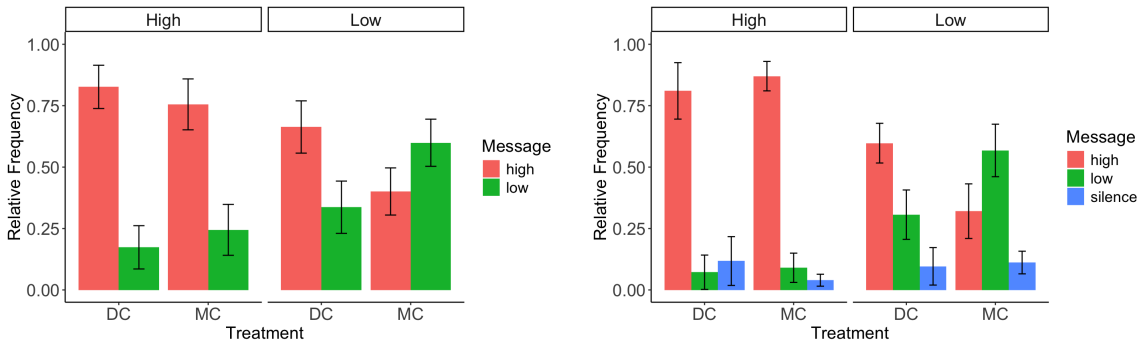
The excluded (default) category is the $q = 1/3$ parametrization, and the first of the two treatments in the session (DC or MC). When looking at different pair types, the default pair is *H-H*. Round refers to the round number within the treatment. Standard errors are clustered at the session level.

Table 9: Peace by treatments.

B.5.4 Disallowing silent messages

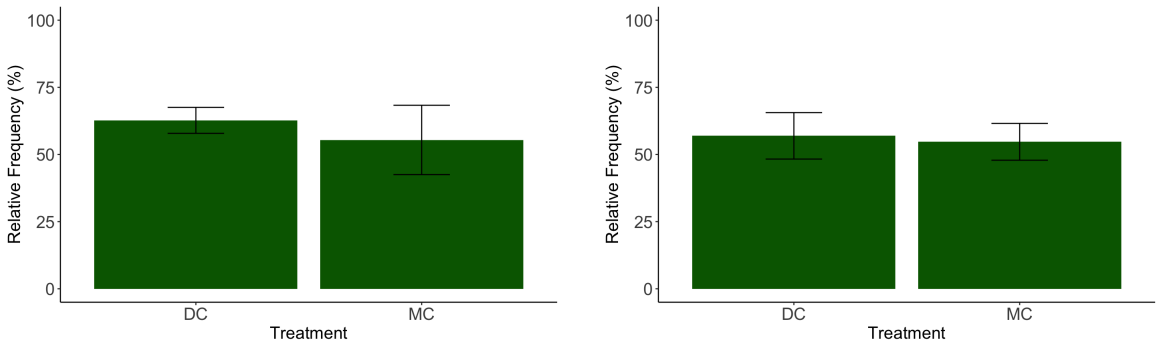
The option of sending a silent message does complicate the theoretical analysis, but does not alter equilibrium properties and we hoped it would make the subjects' task less intimidating. What happens when such an option is not offered?

We ran 4 auxiliary sessions without Silence, for $q = 1/2$ only.[58]



(a) No Silence; $q = 1/2$ (b) With Silence; $q = 1/2$

Figure 16: Sincerity: DC and MC, No Silence v/s With Silence



(a) No Silence; $q = 1/2$ (b) With Silence; $q = 1/2$

Figure 17: Peace: DC and MC, No Silence v/s With Silence

As Figures 16 and 17 show, the results are very closely comparable, which is confirmed in the regression tables in Section B.5.6.

[58] These sessions were run to evaluate the robustness of the results to various possible changes in the design. In addition to having no Silence, we substituted 20 rounds of the simpler No Communication game for the central treatment with a human (participant) mediator run in the original sessions. Two sessions had order DC, NC, MC and two had order MC, NC, DC; 20 rounds for each treatment. The 10 initial introductory rounds were NCI, as in the original data.

B.5.5 Demand strategies in the No-Communication (NC) rounds

In Section 8, we reported the observed frequency of peace in experimental rounds in which subjects played the demand game with no communication (NC). Here we discuss the subjects' demand strategies.

Recall that equilibrium demand strategies under DC when messages are uninformative ($\sigma_H = \sigma_L$ and $\tau_H = 1 - \tau_L - \sigma_L$) are also equilibrium demand strategies in NC. Our theoretical predictions in the absence of communication thus correspond to: $\delta_{0.7}(H) = 1$; $\delta_{0.7}(L) = 0.29$ and $\delta_{0.3}(L) = 0.71$ for both $q = 1/2$ and $q = 1/3$ in equilibrium 2, and, for $q = 1/3$ only, $\delta_{0.7}(H) = 1$ and $\delta_{0.5}(L) = 1$ in equilibrium 1.[59]

Whether in the initial training rounds of the original sessions (NCI) or in the auxiliary sessions where NC is treated symmetrically to DC and MC and played over more rounds (NC), realized demands in the absence of communication align qualitatively with equilibrium 1 under $q = 1/3$ but deviate from equilibrium 2 predictions, both for $q = 1/3$ and $q = 1/2$ (as indeed we also see under DC). In the data, L types demand 50 with high frequency; equilibrium 1 (which exists only for $q = 1/2$) has them demand 50 with probability 1; equilibrium 2, with probability 0. In Table 10, we report the observed frequencies of demand strategies in all NC and DC treatments.

We document in Figure 18 that demand patterns in each of NC and DC in the auxiliary sessions remain similar to what they were in the data for the original experiments. In both data sets, subjects appear to internalize their own type more in the absence of communication. Relative to DC, H types play more aggressively in NC, demanding 70 more frequently and 50 less frequently, and L types play less aggressively, demanding 30 more frequently and 50 less frequently. It is not surprising that the final result is a similar frequency of conflict. Recall that theory predicts equal frequencies of peace in NC and DC, regardless of messages, but different demand strategies unless messages under DC are fully disregarded.

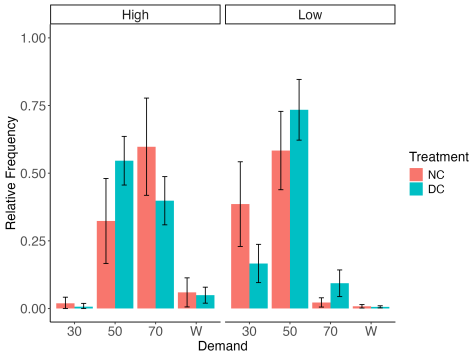
Comparing the data from the new auxiliary sessions to the original data, we see slightly more aggressive behavior for the L types in both treatments in the new sessions, but all relative patterns are unchanged. Here too these auxiliary experiments suggest that the regularities observed in the original data are robust.

[59] For $q = 1/2$, $\delta_{0.7}(H) = 1$; $\delta_{0.7}(L) = 0.29$ and $\delta_{0.3}(L) = 0.71$ is the unique PBE in undominated strategies.

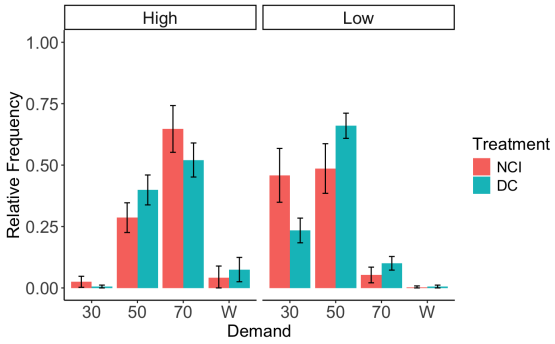
Original sessions								
	$q = \frac{1}{2}$				$q = \frac{1}{3}$			
	0.7	0.5	0.3	w	0.7	0.5	0.3	w
	NCI				NCI			
H	0.65	0.29	0.03	0.04	0.79	0.14	0.01	0.06
L	0.05	0.49	0.46	0.00	0.08	0.66	0.25	0.01
	DC				DC			
H	0.52	0.40	0.01	0.07	0.71	0.17	0.01	0.12
L	0.10	0.66	0.23	0.01	0.18	0.59	0.22	0.01

Auxiliary sessions				
	$q = \frac{1}{2}$			
	0.7	0.5	0.3	w
	NCI			
H	0.55	0.34	0.02	0.08
L	0.01	0.71	0.26	0.02
	NC			
H	0.60	0.32	0.02	0.06
L	0.02	0.58	0.39	0.01
	DC			
H	0.40	0.55	0.01	0.05
L	0.09	0.73	0.17	0.01

Table 10: Demand frequencies in NC and DC.



(a) NC and DC; $q = 1/2$. Auxiliary Sessions.



(b) NCI and DC; $q = 1/2$. Original sessions.

Figure 18: Demands: NC and DC. Auxiliary and original sessions.

B.5.6 Robustness checks regressions

	Dependent variable:		
	Sincerity, $q = 1/2$		
	Original Sessions	No-Silence Sessions	Sessions with Robust MC
MC treatment	0.061	−0.073	−0.024
	(0.062)	(0.068)	(0.061)
Second treatment	0.073	−0.035	0.113*
	(0.061)	(0.059)	(0.061)
L-type	−0.596***	−0.428***	−0.413***
	(0.089)	(0.045)	(0.085)
Round	−0.003	0.002	−0.005**
	(0.002)	(0.002)	(0.002)
MC treatment × L-type	0.198***	0.332***	0.383***
	(0.076)	(0.108)	(0.075)
Second treatment × L-type	0.093	−0.072	−0.111
	(0.075)	(0.094)	(0.077)
Round × L-type	0.004**	−0.001	0.007**
	(0.002)	(0.002)	(0.003)
Constant	0.802***	0.784***	0.705***
	(0.066)	(0.054)	(0.089)
Observations	2,880	1,920	1,920
R^2	0.235	0.155	0.118
Adjusted R^2	0.233	0.152	0.114
Residual Std. Error	0.420 (df = 2872)	0.446 (df = 1912)	0.461 (df = 1912)
Note:			*p<0.1; **p<0.05; ***p<0.01

The excluded (default) categories in the regression are the H-Type, the DC treatment, and the first of the two treatments in the session (DC or MC). Round refers to the round number within the treatment. Standard errors are clustered at the session level.

Table 11: Sincerity, all $q = 1/2$ sessions.

	Dependent variable:	
	Sincerity, $q = 1/3$	
	Original Sessions	Sessions with Robust MC
MC treatment	-0.077	-0.077^{***}
	(0.063)	(0.019)
Second treatment	0.140^{**}	-0.002
	(0.064)	(0.017)
L-type	-0.388^{***}	-0.405^{***}
	(0.145)	(0.083)
Round	-0.003	0.001
	(0.004)	(0.003)
MC treatment $\times$ L-type	0.412^{***}	0.372^{***}
	(0.103)	(0.068)
Second treatment $\times$ L-type	-0.107	-0.045
	(0.105)	(0.070)
Round $\times$ L-type	-0.001	-0.001
	(0.003)	(0.005)
Constant	0.702^{***}	0.812^{***}
	(0.086)	(0.077)
Observations	2,880	1,920
R^2	0.137	0.126
Adjusted R^2	0.135	0.123
Residual Std. Error	0.463 (df = 2872)	0.457 (df = 1912)
Note:		*p<0.1; **p<0.05; ***p<0.01

The excluded (default) categories in the regression are the H-Type, the DC treatment, and the first of the two treatments in the session (DC or MC). Round refers to the round number within the treatment. Standard errors are clustered at the session level.

Table 12: Sincerity, all $q = 1/3$ sessions.

	Dependent variable:		
	Peace, $q = 1/2$		
	Original Sessions	No-Silence Sessions	Sessions with Robust MC
MC treatment	-0.017	-0.073	-0.070
	(0.048)	(0.049)	(0.044)
Second treatment	0.059	-0.007	0.017
	(0.049)	(0.033)	(0.043)
Round	-0.002^{**}	-0.002	0.003
	(0.001)	(0.002)	(0.003)
Pair type *H-L*	0.360^{***}	0.285^{***}	0.351^{***}
	(0.037)	(0.051)	(0.047)
Pair type *L-L*	0.686^{***}	0.563^{***}	0.724^{***}
	(0.052)	(0.050)	(0.041)
Constant	0.213^{***}	0.405^{***}	0.196^{***}
	(0.069)	(0.110)	(0.040)
Observations	1,440	960	960
R^2	0.244	0.179	0.263
Adjusted R^2	0.241	0.175	0.259
Residual Std. Error	0.433 (df = 1434)	0.447 (df = 954)	0.429 (df = 954)
Note:			*p<0.1; **p<0.05; ***p<0.01

The excluded (default) category is the DC treatment, and the first of the two treatments in the session (DC or MC). When looking at different pair types, the default pair is *H-H*. Round refers to the round number within the treatment. Standard errors are clustered at the session level.

Table 13: Peace, all $q = 1/2$ sessions.

	Dependent variable:	
	Peace, $q = 1/3$	
	Original Sessions	Sessions with Robust MC
MC treatment	−0.017	−0.088
	(0.023)	(0.060)
Second treatment	0.063***	−0.063
	(0.023)	(0.061)
Round	0.001	0.001
	(0.003)	(0.004)
Pair type *H-L*	0.313***	0.296***
	(0.026)	(0.030)
Pair type *L-L*	0.723***	0.755***
	(0.031)	(0.017)
Constant	−0.014	0.108***
	(0.051)	(0.032)
Observations	1,440	960
R^2	0.269	0.288
Adjusted R^2	0.266	0.284
Residual Std. Error	0.428 (df = 1434)	0.423 (df = 954)
Note:		*p<0.1; **p<0.05; ***p<0.01

The excluded (default) category is the DC treatment, and the first of the two treatments in the session (DC or MC). When looking at different pair types, the default pair is *H-H*. Round refers to the round number within the treatment. Standard errors are clustered at the session level.

Table 14: Peace, all $q = 1/3$ sessions.

B.6 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: NCI, DC, HMC, MC

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

(Payoffs for HMC: 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]