연세대학교 상경대학

경제연구소

Economic Research Institute

Yonsei Universit

서울시 서대문구 연세로 50

50 Yonsei-ro, Seodaemun-gu, Seoul, Korea

TEL: (+82-2) 2123-4065 FAX: (+82-2) 364-9149

E-mail: yeri4065@yonsei.ac.kr

http://yeri.yonsei.ac.kr/new

Biased Mediation: Selection and Effectiveness

Jin Yeub Kim
Yonsei University

Jong Jae Lee
The Catholic University of Korea

July 2024
2024RWP-230

Economic Research Institute

Yonsei University

Biased Mediation: Selection and Effectiveness*

Jin Yeub Kim† Jong Jae Lee‡

July 20, 2024

Abstract

This paper presents a theory of mediator selection in conflicts that compares biased mediation and unbiased mediation. We determine when and how parties in dispute accept a biased mediator, and characterize optimal mechanisms used by biased mediators when they are selected into mediation in equilibrium. When asymmetric information is significant, parties accept biased mediation as long as the degree of mediator bias is not too high. Biased mediators care more about the payoffs of their favored party. Nevertheless, we find that biased mediators can be equally effective in promoting peace as the unbiased mediator. This implies that, once selected into mediation, the mediator's effectiveness is independent of the degree of mediator bias. The key force of our results is that a biased mediator's optimal recommendation strategies allocate more shares of resource to the favored party while providing a higher chance of a peaceful settlement to the disfavored party.

Key words: Biased Mediation, Conflict, Mechanism Design, Mediator Selection.

JEL codes: C72, D74, D82, F51.

*Lee is deeply grateful to Ying Chen, Edi Karni, Ali Khan, and John Quah for their valuable comments and discussion. We also thank Dongkyu Chang, Liuchun Deng, Johannes Hörner, Kyungmin Kang, Andrés Salamanca, and the participants at Johns Hopkins University seminar and the 2022 Asian Meeting of the Econometric Society (Shenzhen, China) for their helpful feedback.

†School of Economics, Yonsei University, Seoul 03722, Republic of Korea.

‡Corresponding Author, Department of Economics, The Catholic University of Korea, Bucheon 14662, Republic of Korea. Phone: +82 2 2164 4572, Email: piuslee83@gmail.com

1 Introduction

Mediation is the most commonly used method of third-party intervention in conflicts, ranging from labor-management disputes to legal disputes to international conflicts. A natural question is what makes for effective mediators in reducing conflicts. In principle, impartiality is often assumed to be one of the key attributes of mediators for effective mediation.[1] Yet, mediators that arise in practice are often biased in favor of one party and still able to deliver successful mediation outcomes.[2]

In this paper, we provide the theoretical underpinnings for the incidence, function, and effectiveness of biased mediators selected by disputants into mediation. That is, we ask: When can biased mediators be acceptable to both disputants? What kinds of recommendation strategies do biased mediators use? Are biased mediators as effective as unbiased ones at preventing conflict? To answer those questions, we characterize optimal *biased* mediation by applying mechanism design techniques, and determine when and how biased mediators can be accepted by disputants. We then derive novel results comparing the performance of optimal biased mediation and unbiased mediation.

Our formal analysis begins by building a model of mediator selection in a stylized conflict environment á la Hörner, Morelli and Squintani (2015) (hereafter HMS). Two players compete for a resource of fixed size, which can be divided peacefully or else a costly conflict occurs that shrinks the resource. Each player can be either a high or a low type, which is private information that affects how the resource is divided in the event of conflict. We consider a two-stage process: Two players decide whether to accept or reject a mediator with a given degree of bias towards one player. The given mediator maximizes the weighted average of the players' payoffs under mediation, where the weights represent her degree of bias towards

[1] Mediator impartiality is identified as a key attribute of a mediation process in the United Nations Guidance for Effective Mediation published in 2012, available at https://peacemaker.un.org/resources/mediation-guidance; and in the Model Standards of Conduct for Mediators revised in 2005 by the American Arbitration Association, the American Bar Association, and the Association for Conflict Resolution, available at https://www.americanbar.org/groups/dispute_resolution/policy_standards. In many casebooks and textbooks, mediators are described to be impartial, neutral, or disinterested (Izumi, 2010).

[2] See Kydd (2003) for an example of successful biased mediation in an international context. There is plentiful anecdotal and empirical evidence from a variety of contexts that mediators are biased. A brief review of the academic literature on biased mediation is deferred to later in this introduction.

one player over the other.[3] If both players accept, then the mediator collects information privately from the players and recommends a division of the resource with some probability; otherwise conflict occurs. If either player rejects, then the players play the default game in which they simultaneously choose whether to agree to an equal split of the resource, and conflict occurs unless both agree.

We derive the optimal mechanisms used by biased mediators with different degrees of bias, and show how they differ from the optimal mechanism used by the unbiased mediator. We find that a biased mediator uses recommendation strategies that allocate more resources to her favored party while giving the opponent more chances to enjoy a peaceful outcome. This differential treatment arises because the biased mediator must give "information rents" to the disfavored party to incentivize truth-telling. To do so, biased mediators may raise the probabilities of recommending peaceful splits or increase the share allocated to the disfavored party. While the former way benefits both players by saving the resource that might have been wasted in conflict, the latter way only harms the favored party. Thus biased mediators choose the cheaper way of allocating more chances of peace to the disfavored party.

In our mediator selection game, there is always an equilibrium in which the unbiased mediator is accepted with probability one. When the chance of facing a high type is high, unanimous acceptance of a biased mediator can be supported as an equilibrium as long as the degree of bias is not too high. As the players' type become more likely to be high (so that conflict is more likely in the default game), the unanimous acceptance of a more strongly biased mediator can be an equilibrium outcome. This result holds because while a low-type disfavored party has a stronger incentive to pretend to be a high type, the mediator with a higher level of bias can circumvent the incentive compatibility constraint by recommending peaceful split more often to a low-type disfavored party. As mediation is now more attractive, the disfavored party would be willing to accept the mediator who is more biased against him.

Importantly, we establish that biased mediators, if selected into mediation in equilibrium, are exactly as effective in promoting peace as the unbiased mediator. This result is surprising because biased mediators care relatively less about promoting peace than does the unbiased

[3]We use the female pronoun for the mediator, and the male pronoun for the players.

mediator and yet, once they are selected, the effectiveness of mediation is independent of mediator bias. The key intuition is that, in a subset of our parameter space, a biased mediator uses recommendation strategies that treat the low types of two parties asymmetrically in providing information rents: a higher chance of peaceful split to the low type of disfavored party and a higher share of resource to the low type of favored party. By doing so, biased mediators can achieve the same ex ante chance of peace as an unbiased mediator would.

HMS, like us, apply the tools of mechanism design to the study of conflict resolution institutions in international relations.[4] They derive a surprising result that a mediator without enforcement power can be equally effective in promoting peace as an arbitrator with enforcement power. For their results, it is natural to assume, in line with the standard mechanism design literature, that mediators/arbitrators are *unbiased* with the objective of maximizing the ex ante probability of peace. We extend HMS by relaxing their assumption that mediators are unbiased. While unbiased mediators would treat two players symmetrically, biased mediators treat them differently. This leads to technical complications in finding binding incentive constraints at the optimum particularly for a mediator's favored side. Our paper deals with this complication, nicely complementing the work of HMS by further investigating conflict resolution by biased mediators with mechanism design tools.

Our second contribution is to study the effect of mediator bias on mediation success. The literature takes opposing views on whether and how mediator bias influences the effectiveness of mediation. Some scholars argue that mediators are more effective at resolving disputes if they are unbiased or impartial (Beber, 2012; Fisher, 1995; those mentioned in Kleiboer, 1996; Young, 1967). Many other scholars, however, contend that biased mediation can be and often is effective (Kydd, 2003; Rauchhaus, 2006; Savun, 2008; Smith, 1985; Svensson, 2007, 2013).[5] Our result that biased mediators are equally effective as their unbiased counterpart supports the latter view on mediator bias.[6]

[4] A few other papers also apply mechanism design tools to international conflicts, e.g., Bester and Wärneryd (2006), Fey and Ramsay (2009, 2010).

[5] See also Ambrus et al. (2013) who show that as an information intermediary, a biased legislative committee can improve information transmission.

[6] See Salamanca (2024) who studies the effectiveness of mediation with regard to the role of mediator bias in a game-theoretic model of cheap talk.

Our paper is not the first to explicitly include mediator bias in a theoretical model. For example, Kydd (2003) studies a model in which a biased mediator first communicates with one player about the other player's resolve, after which the two players bargain with each other; unlike us, such a mediator only has a role of an advisor who transmits information. Importantly, Beber (2012) points out that most theoretical and empirical analyses study the initiation of mediation separately from its effectiveness. Beber (2012) argues, unlike us, that biased mediators are relatively ineffective; and suggests a counterintuitive selection effect that such mediators are more likely to select themselves into mediation, focusing on the incentives of the mediator to engage in mediation. The novelty of our paper is in linking the effectiveness of biased mediator to the initiation of mediation focusing on the incentives of disputants to seek the assistance of a mediator. By dealing with how the mediation process starts, our analysis implies that only those mediators who are effective at promoting peace, whether biased or unbiased, are more likely to be selected by disputants into mediation.

Methodologically, we apply Celik and Peters' (2011) approach to model a mediator selection game.[7] In particular, we augment the baseline conflict model of HMS to include a stage where the players choose whether to accept a biased mediator and play the mediation game, or to resort to the default game of conflict.[8] By doing so, our analysis accounts for the strategic considerations of disputants during both the selection and the implementation stages of mediation. Celik and Peters (2011) study rejection on the equilibrium path and establish a condition under which all the implementable outcomes are truthfully implementable. In the context of mediation, both disputants must consent to the involvement of a mediator as a third party; so our interest is in identifying biased mediators that both disputants accept. Hence, we are only concerned about unanimous acceptance of a biased mediator in equilibrium.

As in HMS, we use the international conflict context to describe the model and results. In international conflicts, adversaries have repeatedly turned to the aid of mediation by a third

[7]See also Cramton and Palfrey (1995) and Tan and Yilankaya (2007) for the general mechanism design problems that involve agreements about how to play the default game.

[8]This extension was mentioned in the 2010 working paper version of HMS; which is relevant to our framework because we consider the question of when biased mediators are accepted.

party to help them reduce the risk of costly conflicts or reach a negotiated settlement.[9] Our model can be applied to other kinds of conflict situations that involve mediation. For example, in labor-management disputes, a union and an employer utilize the services of mediators to settle on terms of a collective bargaining agreement.[10] In litigation and legal disputes including partnership dissolution, some jurisdictions require court-annexed alternative dispute resolution such as arbitration or mediation for parties to reach pretrial settlements.[11]

This paper is organized as follows. Section 2 introduces our model of mediator selection. Section 3 characterizes optimal mediation. Section 4 characterizes equilibria of mediator selection and compare the effectiveness of biased and unbiased mediation. Section 5 concludes. All the proofs and supplementary materials are in the Appendices.

2 The Model

We present a model of mediator selection in a conflict environment as a two-stage game. In the first stage, two players with private information simultaneously choose whether to accept a mediator. In the second stage, if the mediator is unanimously accepted, then its optimal mechanism is implemented; otherwise the default game as an outside option is played.

2.1 The Baseline Environment

We begin by describing a stylised bilateral conflict environment with asymmetric information á la HMS. Two players (1 and 2) compete for a given resource of size normalized to one. If two players agree to a peaceful sharing of the resource, then each receives the agreed share. If not, an outright conflict (say, war) occurs and shrinks the value of the resource to $\theta < 1$. Each player can be of high (H) or low (L) type, privately and independently drawn from the same distribution with probability $q \in (0, 1)$ and $(1 - q)$ respectively. The player's type

[9]See Bercovitch and Gartner (2008, 5) and Frazier and Dixon (2006, 395). Although the number varies across the databases, the incidence of the third-party mediation accounts for about a 30 to 40 percent rate.

[10]See https://www.fmcs.gov/services/resolving-labor-management-disputes.

[11]For different legal cases in which mediation can be used, see the Mediation section of the online resources provided by the American Bar Association, available at https://www.americanbar.org/groups/public_education/resources/law_related_education_network/how_courts_work.

can be thought of as his resolve, potential, or strength in conflict, determining the expected war payoffs. When the two players' types are equal, each receives $\theta/2$. When one player is H and the other is L, the H-type player's expected share in war is $p > 1/2$ so his expected war payoff is $p\theta > 1/2$ while the L-type player receives $(1-p)\theta < 1/2$.

In describing our results, we often use the following two parameters, as used in HMS:

$$\lambda \equiv \frac{q}{1-q} \quad \text{and} \quad \gamma \equiv \frac{p\theta - 1/2}{1/2 - \theta/2}.$$

The parameter $\lambda > 0$ is the H/L type odds ratio; and $\gamma > 0$ measures the H-type's gain from war against an L-type instead of accepting the equal split relative to its loss from war against an H-type instead of accepting the equal split. If $\lambda \geq \gamma$ (i.e., $q\theta/2 + (1-q)p\theta \leq 1/2$), then war can always be averted by the equal split. So, as in HMS, we assume $\lambda < \gamma$ throughout.

2.2 The Default Game

We consider a simple default game where the two players simultaneously announce "peace" or "war." If they both announce peace, they equally share the resource (i.e., each player receives 1/2), otherwise war takes place.[12] Let $a_i \in \{P, W\}$ denote an announcement by player i, where P denotes peace and W denotes war.

In equilibrium, each type of each player chooses peace or war to maximize its expected payoff given the beliefs in the default game. We restrict attention to the equilibrium that maximizes the ex ante probability of peace, referred to as the peace-maximizing equilibrium. We will allow for the default game to be played under updated beliefs. Let $\{q_1, q_2\}$ represent an arbitrary belief system, where each q_i gives the probability that player $-i$ is attributing to the event that player i is type H when the players are playing the default game. Under the arbitrary belief system $\{q_1, q_2\}$, the peace-maximizing equilibrium of the default game is characterized as follows: When $q_1 \leq (1-q_1)\gamma$ or $q_2 \leq (1-q_2)\gamma$, H-types announce war

[12]Our default game is a simplification of the "war declaration game" described in the 2010 working paper version of HMS. More generally, we can think of there being a default game for any proposed split $(x, 1-x)$, where $x \in [0,1]$. If both players announce peace, then the proposed split is accepted and implemented.

and L-types announce peace; when $q_1 > (1 - q_1)\gamma$ and $q_2 > (1 - q_2)\gamma$, both types announce peace. The detailed derivation is presented in Appendix A.

2.3 The Mediation Game

The two players can use mediation. By invoking the revelation principle, we can set up the mediation game as a direct-revelation mechanism, without loss of generality. We assume that such a mechanism is offered by a third party, which we will call the *mediator*. In the mediation game, each player i privately sends a report m_i to the mediator, if selected into mediation. Given reports $m = (m_1, m_2)$, the mediator publicly recommends a peaceful split of the resource $(b_m, 1 - b_m)$ with probability p_m. We assume that the mediator's recommendations are implemented.[13] With probability $1 - p_m$, the mediation fails and the players go to war. We refer to b_m as the share of the resource allocated to player 1 and p_m as the probability of recommending peaceful split, given the reported type profile m.

In equilibrium, each type of each player decides whether to participate or not and chooses a report to maximize its expected payoff, given the beliefs and the mechanism offered by the mediator in the mediation game. Again by the revelation principle, without loss of generality, we will focus on equilibria in which the players agree to participate in mediation and truthfully report their types. Our aim in this paper is to study whether a biased mediator can be accepted by the players, rather than whether the given mediator's mechanism can be truthfully implemented. So unlike Celik and Peters (2011), we delineate the mediator's preferences and its specific objective.

We assume that any given mediator has no private information and does not know the players' types but holds some beliefs about the players' types. What is important in our framework is that the mediator may be *biased*, represented by her objective. We specify the mediator's objective to be the maximization of the weighted average of the players' payoffs

[13]To be precise, a mediator in our setting is assumed to have enforcement power. While enforcement may not be feasible in some situations, HMS show that a mediator without enforcement power is exactly as effective as a mediator with enforcement power. Our modeling approach is a shortcut to describe mediation, abstracting away from the discussion about enforcement, which allows us to greatly simplify the dimensionality of our analysis.

under mediation, where the weights are defined by her bias towards one player over the other. That is, the mediator's objective function is defined as

$$w_\beta \equiv \beta U_1 + (1-\beta)U_2,$$

where U_1 and U_2 are payoffs for players 1 and 2 respectively under mediation, and $\beta \in [0,1]$ measures the degree by which the mediator is biased towards player 1. If $\beta > 1/2$ (resp. $\beta < 1/2$), the mediator is said to be biased in favor of player 1 (resp. player 2); and if $\beta = 1/2$, the mediator is unbiased. Any mediator given to the players can be identified by its degree of bias β. We use β-mediator to refer to the mediator with bias β.

Under the arbitrary belief system $\{q_1, q_2\}$, the β-mediator mediates according to a specified mediation plan (or mechanism) that maximizes her expected payoff given beliefs $\{q_1, q_2\}$ subject to incentive compatibility and participation constraints. We assume that the mediator commits to her mechanism, denoted by $\mathcal{M}(\beta, q_1, q_2)$. With a slight abuse of notation, we write $m = m_1 m_2$ instead of (m_1, m_2), and let $M = \{HH, HL, LH, LL\}$ denote the set of all possible message-pairs m.[14] Thus β-mediator with posteriors $\{q_1, q_2\}$ can be represented by her mechanism $\mathcal{M}(\beta, q_1, q_2) = (p_m, b_m)_{m \in M}$.

2.4 Mediator Selection Stage

Nature randomly chooses a mediator bias β from $[0,1]$. The β-mediator offers to mediate the dispute. Each player (already knowing his own private type) simultaneously and independently chooses whether to accept or reject the β-mediator. Because we are primarily interested in cases where biased mediators are accepted with probability one, we focus on pure strategies. Formally, a selection strategy of player i of type $t_i \in T_i = \{H, L\}$ is $v_i : T_i \to \{1, 0\}$, where 1 denotes "accept" and 0 denotes "reject."

We can represent player i's selection decision with the binary variable s_i (equals to y if player i accepts, to n otherwise). We refer to $s = (s_1, s_2)$ as a selection profile. After the selection stage, both players and the mediator observe the selection decisions made by

[14]Because we focus on direct mechanisms, we write the message space as same as the type space.

each player. This observation allows the players *and the mediator* to update beliefs on the players' types. Let q_i^y denote the posterior belief probability on player i being H-type if player i accepts; and q_i^n denote the posterior belief probability on player i being H-type if player i rejects. The beliefs are common knowledge among the players and the mediator.

If both players accept unanimously (i.e., $s = (y, y)$), then the game moves to the mediation game. Otherwise, the players play the default game. There are three different selection profiles under which the mediator will be rejected by at least one of the players. We let $R = \{(y, n), (n, y), (n, n)\}$ denote the set of these selection profiles.

2.5 The Equilibrium

Our solution concept is Perfect Bayesian equilibrium. An equilibrium consists of a selection strategy profile $\{v_1^*, v_2^*\}$, a mediation game mechanism $\mathcal{M}(\beta, \cdot, \cdot)$ offered by the given β-mediator, default game strategies $\{a_1^{s*}, a_2^{s*}\}_{s\in R}$, and beliefs $\{q_i^{y*}, q_i^{n*}\}_{i=1,2}$, which together satisfy the following conditions:

(i) $\mathcal{M}(\beta, q_1^{y*}, q_2^{y*})$ maximizes the β-mediator's expected payoff given her bias β subject to incentive compatibility and participation constraints under the belief system $\{q_1^{y*}, q_2^{y*}\}$.

(ii) For each selection profile $s \in R$ leading to the default game, $\{a_1^{s*}, a_2^{s*}\}$ constitutes the peace-maximizing equilibrium of the default game under the belief system $\{q_1^{s_1*}, q_2^{s_2*}\}$.

(iii) For each player i and each type t_i, $v_i^*(t_i)$ maximizes the expected continuation payoff, given player $-i$'s selection strategy v_{-i}^*, $\mathcal{M}(\beta, q_1^{y*}, q_2^{y*})$, and $\{a_1^*, a_2^*\}$.

(iv) $\{q_i^{y*}\}_{i=1,2}$ and $\{q_i^{n*}\}_{i=1,2}$ are updated by the Bayes rule on the equilibrium path. That is, $q_i^{y*} = \frac{qv_i^*(H)}{qv_i^*(H)+(1-q)v_i^*(L)}$ if v_i^* is not constant at 0 and $q_i^{n*} = \frac{q[1-v_i^*(H)]}{q[1-v_i^*(H)]+(1-q)[1-v_i^*(L)]}$ if v_i^* is not constant at 1.

Condition (i) restricts a Bayesian Nash equilibrium of the mediation game to the one induced by a truthfully implementable mechanism. So when both players accept the β-mediator, then their equilibrium mediation payoffs are determined by the optimal mechanism $\mathcal{M}(\beta, q_1^{y*}, q_2^{y*})$ that is incentive compatible and individually rational. In Section 3, we formulate the optimal mechanism $\mathcal{M}(\beta, q_1^{y*}, q_2^{y*})$ in the mediation game.

Condition (ii) dictates that when either player rejects mediation, after learning the realized selection profile $s \in R$, each player chooses to announce peace or war according to the peace-maximizing equilibrium of the default game under the belief system $\{q_1^{s_1*}, q_2^{s_2*}\}$. Note that there is no consistency requirement for the off the equilibrium path rejection beliefs. It is not a concern, as our interest is to discuss whether a biased mediator *can* be accepted unanimously. Also, we are not concerned with characterizing all the implementable mechanisms by the mediator (including the ones feasible only through equilibrium rejection); so we do not consider rejections on the equilibrium path.[15]

3 Biased Mediator's Optimal Mechanism

In this section, we characterize the β-mediator's optimal mechanism used in the mediation game, after it has been unanimously accepted by both types of both players.

3.1 Preliminaries

For any given mechanism $\mathcal{M}(\beta, q_1, q_2)$ under arbitrary beliefs $\{q_1, q_2\}$, the incentive compatibility constraints for player 1 of each type are characterized as:

$$\begin{aligned}
(IC_{1H}) \quad & q_2 \left[p_{HH}b_{HH} + (1 - p_{HH})(\theta/2)\right] + (1 - q_2) \left[p_{HL}b_{HL} + (1 - p_{HL})p\theta\right] \\
& \geq q_2 \left[p_{LH}b_{LH} + (1 - p_{LH})(\theta/2)\right] + (1 - q_2) \left[p_{LL}b_{LL} + (1 - p_{LL})p\theta\right]; \\
(IC_{1L}) \quad & q_2 \left[p_{LH}b_{LH} + (1 - p_{LH})(1 - p)\theta\right] + (1 - q_2) \left[p_{LL}b_{LL} + (1 - p_{LL})(\theta/2)\right] \\
& \geq q_2 \left[p_{HH}b_{HH} + (1 - p_{HH})(1 - p)\theta\right] + (1 - q_2) \left[p_{HL}b_{HL} + (1 - p_{HL})(\theta/2)\right].
\end{aligned}$$

The left-hand-sides are the expected payoffs for player 1 of H-type and L-type, respectively, under beliefs $\{q_1, q_2\}$ given that both players report their types honestly in $\mathcal{M}(\beta, q_1, q_2)$. The right-hand-sides are the expected payoffs for player 1 of each of the two types, if he lies about

[15] In fact, in our setting considering rejections on the equilibrium path does not extend the set of feasible outcomes. We briefly discuss the relevant analysis in Remark 2.

his type in $\mathcal{M}(\beta, q_1, q_2)$ while player 2 remains honest. Similarly, the incentive compatibility constraints for player 2 of each type are characterized as:

$$
\begin{aligned}
(IC_{2H}) \quad & q_1 \left[p_{HH}(1-b_{HH}) + (1-p_{HH})(\theta/2)\right] + (1-q_1)\left[p_{LH}(1-b_{LH}) + (1-p_{LH})p\theta\right] \\
& \geq q_1 \left[p_{HL}(1-b_{HL}) + (1-p_{HL})(\theta/2)\right] + (1-q_1)\left[p_{LL}(1-b_{LL}) + (1-p_{LL})p\theta\right]; \\
(IC_{2L}) \quad & q_1 \left[p_{HL}(1-b_{HL}) + (1-p_{HL})(1-p)\theta\right] + (1-q_1)\left[p_{LL}(1-b_{LL}) + (1-p_{LL})(\theta/2)\right] \\
& \geq q_1 \left[p_{HH}(1-b_{HH}) + (1-p_{HH})(1-p)\theta\right] + (1-q_1)\left[p_{LH}(1-b_{LH}) + (1-p_{LH})(\theta/2)\right].
\end{aligned}
$$

Because war can be initiated unilaterally, the players should be willing to participate in the mediation process after they have accepted the mediator. The participation constraints for players 1 and 2 of each type, given mechanism $\mathcal{M}(\beta, q_1, q_2)$, are characterized as:

$$
\begin{aligned}
(PC_{1H}) \quad & q_2 \left[p_{HH}b_{HH} + (1-p_{HH})(\theta/2)\right] + (1-q_2)\left[p_{HL}b_{HL} + (1-p_{HL})p\theta\right] \\
& \geq q_2(\theta/2) + (1-q_2)p\theta, \\
(PC_{1L}) \quad & q_2 \left[p_{LH}b_{LH} + (1-p_{LH})(1-p)\theta\right] + (1-q_2)\left[p_{LL}b_{LL} + (1-p_{LL})(\theta/2)\right] \\
& \geq q_2(1-p)\theta + (1-q_2)(\theta/2); \\
(PC_{2H}) \quad & q_1 \left[p_{HH}(1-b_{HH}) + (1-p_{HH})(\theta/2)\right] + (1-q_1)\left[p_{LH}(1-b_{LH}) + (1-p_{LH})p\theta\right] \\
& \geq q_1(\theta/2) + (1-q_1)p\theta, \\
(PC_{2L}) \quad & q_1 \left[p_{HL}(1-b_{HL}) + (1-p_{HL})(1-p)\theta\right] + (1-q_1)\left[p_{LL}(1-b_{LL}) + (1-p_{LL})(\theta/2)\right] \\
& \geq q_1(1-p)\theta + (1-q_1)(\theta/2),
\end{aligned}
$$

where the right-hand-sides are the expected payoffs in war.

The β-mediator's expected payoff is the (ex ante) expected value of the β-weighted average of the two players' payoffs in $\mathcal{M}(\beta, q_1, q_2)$ given the beliefs about the players' types $\mathbf{q} \equiv (q_1, q_2)$, which is formulated as follows:

$$
W_\beta\left(\mathcal{M}(\beta, q_1, q_2)\right) \equiv E_{\mathbf{q}}\Big[\beta\big(p_t b_t + (1-p_t)d_t\big) + (1-\beta)\big(p_t(1-b_t) + (1-p_t)(\theta - d_t)\big)\Big],
$$

where b_t is the share of the resource allocated to player 1 under the peaceful split $(b_t, 1-b_t)$

recommended with probability p_t, and d_t is the share of the remaining resource allocated to player 1 under war, given the reported type profile $t \in T = M$.[16]

We find it useful to decompose $W_\beta\left(\mathcal{M}(\beta, q_1, q_2)\right)$ into three components, as follows:

$$\begin{aligned} W_\beta\left(\mathcal{M}(\beta, q_1, q_2)\right) =& (2\beta - 1)E_{\mathbf{q}}\big[p_t(b_t - d_t)\big] + (1-\beta)(1-\theta)E_{\mathbf{q}}\big[p_t\big] + E_{\mathbf{q}}\big[D_t\big] \\ \text{or, equivalently, } =& (1-2\beta)E_{\mathbf{q}}\big[p_t\big((1-b_t) - (\theta - d_t)\big)\big] + \beta(1-\theta)E_{\mathbf{q}}\big[p_t\big] + E_{\mathbf{q}}\big[D_t\big], \end{aligned}$$

where $D_t \equiv \beta d_t + (1-\beta)(\theta - d_t)$ for each $t \in T$. The $E_{\mathbf{q}}\big[p_t(b_t - d_t)\big]$ or $E_{\mathbf{q}}\big[p_t\big((1-b_t)-(\theta-d_t)\big)\big]$ in the first term is the ex-ante expected gain of player 1 or 2 from the recommended split relative to his war payoff, and $E_{\mathbf{q}}\big[p_t\big]$ in the second term is the ex ante expected probability of peace. The last term, $E_{\mathbf{q}}\big[D_t\big]$, is a constant given β. We can easily see that if $\beta = 1$, then the mediator's preference is fully aligned with player 1's ex ante gain; and if $\beta = 0$, then it is fully aligned with player 2's. If a mediator is unbiased, i.e., $\beta = 1/2$, then she cares only about maximizing the ex ante probability of peace.

Mediator's Optimal Mediation Programme The optimal mediation programme **(P)** determines (p_t, b_t) for each $t \in T$ to maximize the expected payoff of β-mediator under the beliefs $\{q_1, q_2\}$:[17]

$$\textbf{(P)} \qquad \max_{(p_t, b_t)_{t\in T}} W_\beta\left((p_t, b_t)_{t\in T}\right)$$

$$\text{subject to } (IC_{it_i}) \text{ and } (PC_{it_i}), \forall i, \forall t_i;\ 0 \le p_t \le 1 \text{ and } 0 \le b_t \le 1, \forall t \in T.$$

3.2 Optimal Mechanism

We focus on unanimous acceptance by all types of all players, i.e., $v_i(t_i) = 1$ for all t_i and all i. Given any β-mediator, if the players follow such a strategy, then no learning occurs. Because nothing is revealed after the selection stage, the posterior beliefs about the players' types are identical to the prior beliefs: $q_i^{y*} = q$ for all i. Given the beliefs $\mathbf{q}^* = (q, q)$, the

[16]Because $\mathcal{M}(\beta, q_1, q_2)$ is a direct revelation mechanism, we will henceforth use the notation $(p_t, b_t)_{t\in T}$ instead of $(p_m, b_m)_{m\in M}$ where $T = M$.

[17]In Appendix B, we provide an explicit expression of the mediator's expected payoff in terms of a player's payoff as well as a reduced form of the mediator's optimization problem **(P)**.

β-mediator would mediate according to her optimal mechanism $\mathcal{M}(\beta, \mathbf{q}^*)$ that solves **(P)**.

In characterizing optimal mechanisms, we introduce the following parameter. Let du_{it_i} denote the *information rent* that player i of type t_i receives in mediation if all players participate truthfully given the belief $\lambda = \frac{q}{1-q}$, which is defined as:

$$du_{1H} \equiv \lambda p_{HH} B_{HH} + p_{HL} B_{HL}, \qquad du_{2H} \equiv \lambda p_{HH}(1 - B_{HH}) + p_{LH}(1 - B_{LH}),$$
$$du_{1L} \equiv \lambda p_{LH} B_{LH} + p_{LL} B_{LL}, \qquad du_{2L} \equiv \lambda p_{HL}(1 - B_{HL}) + p_{LL}(1 - B_{LL}),$$

where $B_t = \frac{b_t - d_t}{1-\theta}$ for $t \in T$. The B_t is the ratio of player 1's net gain or loss from the mediator's recommended split (relative to war) over the lost resource from war, conditional on participation and truthful type revelation of both players in mediation given a type profile $t \in T$. Hence, du_{it_i} represents the expected net gain (or loss) from mediation instead of waging war. Then the interim expected payoff for player i of type t_i in the mechanism $\mathcal{M}(\beta, \mathbf{q}^*)$ can be written as $U_i(\mathcal{M}|t_i) = du_{it_i}(1-\theta)(1-q) +$ ("i of type t_i's expected war payoff").

We now describe the main features of optimal mechanisms under the beliefs $\mathbf{q}^* = (q, q)$ by reporting the mechanism probabilities $(p_t)_{t \in T}$ and the information rents $(du_{it_i})_{i \in \{1,2\}, t_i \in T_i}$. For the unbiased mediator's optimal mechanism, we can use Lemma 1 of HMS, stated below without proof but with our additional specifications of information rents.

Lemma 1 (Unbiased Mediator's Optimal Mechanism, Lemma 1 of HMS)**.** *Suppose that the mediator is unbiased, $\beta = 1/2$. Then the optimal mechanism satisfies the following:*

(1) The L-type's incentive compatibility constraint and the H-type's participation constraint bind, and the other constraints do not.

(2) For $\lambda \leq \gamma/2$, low-type dyads (L, L) do not fight ($p_{LL} = 1$), asymmetric dyads (H, L) and (L, H) are recommended a peaceful split with probability $p_{HL} = p_{LH} = \frac{1}{1+\gamma-2\lambda} \in (0, 1)$, and high-type dyads (H, H) always fight ($p_{HH} = 0$). The rents are $du_{iH} = 0$ and $du_{iL} = \frac{1+\gamma}{2(1+\gamma-2\lambda)} > 0$ for $i = 1, 2$.

(3) For $\gamma/2 < \lambda < \gamma$, low-type dyads and asymmetric dyads do not fight ($p_{LL} = p_{HL} = p_{LH} = 1$) and high-type dyads fight with positive probability $\big(p_{HH} = \frac{2\lambda-\gamma}{\lambda(1+\gamma-\lambda)} \in (0, 1)\big)$. The rents are $du_{iH} = 0$ and $du_{iL} = \frac{(1+\gamma)(\lambda+1)}{2(1+\gamma-\lambda)} > 0$ for $i = 1, 2$.

Unlike the unbiased mediator, a biased mediator may treat players differently depending on her degree of bias, and so we cannot restrict attention to symmetric recommendations. Accordingly, the derivations for the solution of biased optimal mediation programme are complicated; we relegate the proof to Appendix C. Here we report the main features of the optimal mechanism used by a biased mediator with $\beta > 1/2$. The analogous results for the case of $\beta < 1/2$ can be stated by interchanging the roles of players 1 and 2.

Proposition 1 (Biased Mediator's Optimal Mechanism)**.** *Suppose that the mediator is biased in favor of player 1, $\beta > 1/2$. Let $\hat{\beta} \equiv \frac{1+\gamma}{2(1+\gamma-\lambda)}$. Then the optimal mechanism satisfies the following:*

(1) When the mediator is strongly biased towards player 1, i.e., $\beta > \hat{\beta}$, the L-type incentive compatibility constraints (IC_{1L}) and (IC_{2L}), and the participation constraints for player 2 (PC_{2H}) and (PC_{2L}) bind.

(a) For $\gamma > 1$ and $\lambda \leq (\gamma-1)/2$, low-type dyads do not fight ($p_{LL} = 1$), (H, L)-dyads fight with positive probability $\left(p_{HL} = \frac{1}{(1+\gamma)/2-\lambda} \in (0, 1]\right)$; otherwise they always fight ($p_{LH} = p_{HH} = 0$). The H-type participation constraint for player 1 (PC_{1H}) additionally binds. The rents are $du_{1H} = du_{2H} = du_{2L} = 0$, and $du_{1L} = \frac{1+\gamma}{2((1+\gamma)/2-\lambda)}$.

(b) For $(\gamma-1)/2 < \lambda < \gamma$, the players do not fight when player 2 is of L-type ($p_{LL} = p_{HL} = 1$); otherwise they always fight ($p_{HH} = p_{LH} = 0$). The H-type incentive compatibility constraint for player 1 (IC_{1H}) additionally binds. The rents are $du_{1H} = 1 - \frac{1+\gamma}{2(1+\lambda)} > 0$, $du_{1L} = 1 + \frac{\lambda(1+\gamma)}{2(1+\lambda)}$, and $du_{2H} = du_{2L} = 0$.

(2) When the mediator is moderately biased towards player 1, i.e., $1/2 < \beta \leq \hat{\beta}$, the L-type incentive compatibility constraints (IC_{1L}) and (IC_{2L}), and the H-type participation constraints (PC_{1H}) and (PC_{2H}) bind.

(a) For $\gamma > 1$ and $\lambda \leq (\gamma-1)/2$, low-type dyads do not fight ($p_{LL} = 1$), (H, L)-dyads fight with positive probability $\left(p_{HL} = \frac{1}{(1+\gamma)/2-\lambda} \in (0, 1]\right)$; otherwise they always fight ($p_{LH} = p_{HH} = 0$). The L-type participation constraint for player

2 (PC_{2L}) additionally binds. The rents are $du_{1H} = du_{2H} = du_{2L} = 0$, *and* $du_{1L} = \frac{1+\gamma}{2((1+\gamma)/2-\lambda)}$.

(b) For $(\gamma-1)/2 < \lambda \leq \gamma/2$, *the players do not fight when player 2 is of L-type* ($p_{LL} = p_{HL} = 1$), *high-type dyads always fight* ($p_{HH} = 0$), *and* (L,H)*-dyad fights with positive probability* $\big(p_{LH} = \frac{1-\gamma+2\lambda}{1+\gamma-2\lambda} \in (0,1)\big)$. *The rents are* $du_{1H} = du_{2H} = 0$, $du_{1L} = \frac{1+\gamma}{2}$, *and* $du_{2L} = \frac{(1+\gamma)(1-\gamma+2\lambda)}{2(1+\gamma-2\lambda)} < du_{1L}$.

(c) For $\gamma/2 < \lambda < \gamma$, *low-type dyads and asymmetric dyads do not fight* ($p_{LL} = p_{HL} = p_{LH} = 1$) *and high-type dyads fight with positive probability* $\big(p_{HH} = \frac{2\lambda-\gamma}{\lambda(1+\gamma-\lambda)} \in (0,1)\big)$. *The rents are* $du_{iH} = 0$ *and* $= du_{iL} = \frac{(1+\gamma)(\lambda+1)}{2(1+\gamma-\lambda)}$ *for* $i = 1, 2$.

Direct comparison of Lemma 1 and Proposition 1 shows that the optimal mechanisms of moderately biased mediators ($1/2 < \beta \leq \hat{\beta}$) and of the unbiased mediator ($\beta = 1/2$) coincide when the players' type is likely high ($\gamma/2 < \lambda < \gamma$). But when the players' type is likely low ($\lambda \leq \gamma/2$) or when the mediator is strongly biased ($\beta > \hat{\beta}$), the biased mediator's optimal mechanism is significantly more complex than the unbiased mediator's counterpart in terms of the probabilities of recommending a peaceful split to asymmetric dyads and the rents for L-type of two players. In particular, the unbiased mediator recommends a peaceful settlement to asymmetric dyads (L,H) and (H,L) with the same probability, essentially treating the L-types of two players equally and allocating the same rents; however, the biased mediator allocates more information rents to the L-type of the favored player while allowing the L-type of the disfavored player enjoy peaceful settlement more often.

The key step in the intuition for this result is realizing that to keep in check the incentive compatibility of L-type of a disfavored player (say, player 2), the mediator in favor of player 1 can either raise the probability of recommending peaceful split p_{HL} or increase the share $1 - b_{HL}$ allocated to the L-type player 2 in (H,L) dyads. The former benefits an H-type of player 1 while the latter only harms both types of player 1. Thus the biased mediator chooses the "cheaper" way, offering peaceful settlement more often to the L-type of disfavored player than to that of favored player and providing more information rents to the L-type of favored player than to that of disfavored player. When the mediator is strongly biased towards player 1 ($\beta > \hat{\beta}$), this effect is aggravated for $(\gamma - 1)/2 < \lambda < \gamma$: the strongly biased mediator never triggers war when the disfavored player is of L-type while always triggering war otherwise; further, both types of the favored player receive positive information rents, whereas the disfavored player receives zero rent regardless of her type.

Remark 1. When a mediator is biased in favor of player 1, then her optimal recommendation strategies exhibit two key features. (1) An L-type player 2 enjoys more chance of peaceful settlement than an L-type player 1: $p_{HL} \geq p_{LH}$. (Note that $p_{LL} = 1$.) (2) An L-type player 1 obtains more information rents than an L-type player 2: $du_{1L} \geq du_{2L}$. This inequality is strict iff the mediator is strongly biased ($\beta > \hat{\beta}$) or the odds ratio is small ($\lambda \leq \gamma/2$). That is, a biased mediator uses recommendation strategies that allocate more shares of the resource to her favored player while assigning higher chances of peace to her disfavored player, compared to the unbiased mediator's optimal recommendation strategies.

4 Mediator Selection and Effectiveness

In this section, we show when and which biased mediators can be selected into mediation in equilibrium, and examine the effectiveness of biased mediators. Proofs are in Appendix D.

4.1 Selection of Biased Mediators

We study equilibria of our mediator selection game where the given β-mediator is accepted with probability one, i.e., $v_i(t_i) = 1$ for all i and t_i. In such an equilibrium, if it exists, a player's acceptance conveys no information.[18] It follows from the Bayes rule that beliefs remain the same as their interim values after observing the acceptance of both players. Hence, the β-mediator's optimal mechanism is computed under prior beliefs, given in Lemma 1 and Proposition 1.

[18]Given any β-mediator, there is always a trivial pooling equilibrium in which $v_i(t_i) = 0$ for all i and t_i. This is an equilibrium because, as long as the other player is expected to reject, each player must expect that his acceptance cannot make any difference, so he might as well reject too even if he really would prefer the given mediator. Further, an equilibrium in which $v_i(t_i) = 1$ for all i and t_i, if it exists, is the only pooling equilibrium where acceptance of a given mediator can occur with positive probability.

The following result shows for what range of values for β there exists an equilibrium in which all types of all players accept the β-mediator (i.e., unanimous acceptance).

Proposition 2. *Let* $\hat{\beta} \equiv \frac{1+\gamma}{2(1+\gamma-\lambda)}$.

(1) If $\gamma > \lambda \geq \frac{\gamma(1+\gamma)}{2(2+\gamma)}$*, then unanimous acceptance of* β*-mediator for any given* $\beta \in [1-\hat{\beta}, \hat{\beta}]$ *constitutes an equilibrium.*

(2) If $\lambda < \frac{\gamma(1+\gamma)}{2(2+\gamma)}$*, then unanimous acceptance of the unbiased mediator* ($\beta = 1/2$) *is an equilibrium; whereas any given biased mediator* ($\beta \neq 1/2$) *is not selected into mediation in equilibrium.*

Figure 1 illustrates Proposition 2. There is always an equilibrium in which the players would agree to accept unanimously the unbiased mediator, if offering to mediate the dispute, regardless of the H/L odds ratio λ. We are interested in whether unanimous acceptance of a *biased* mediator can also be an equilibrium. When the chance of facing a high-type is low such that $\lambda < \frac{\gamma(1+\gamma)}{2(2+\gamma)}$, unanimous acceptance of a biased mediator is never an equilibrium outcome. However, when the chance of facing a high-type is high such that $\lambda \geq \frac{\gamma(1+\gamma)}{2(2+\gamma)}$, unanimous acceptance of a biased mediator can be supported as an equilibrium as long as the degree of bias is not too high ($\beta \leq \hat{\beta}$). The more likely the H-type, the more strongly biased mediator can be unanimously accepted in equilibrium; that is, $\hat{\beta}$ increases in λ.

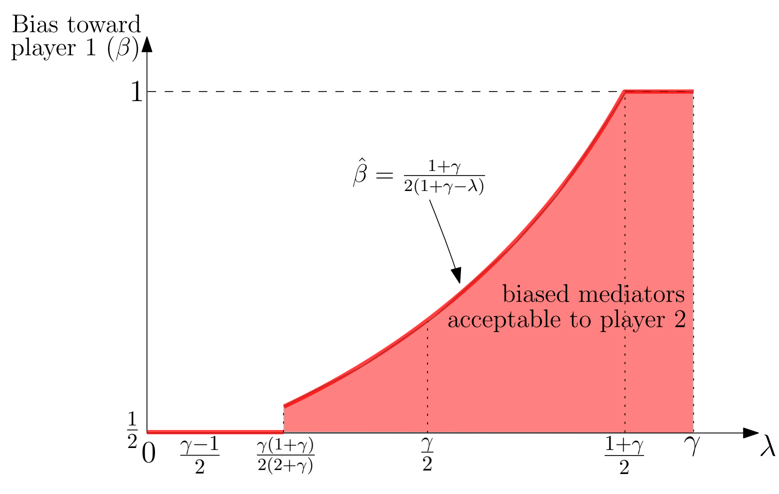


Figure 1: Region where unanimous acceptance of β-mediators with $\beta \geq 1/2$ is an equilibrium. *Note*: In the figure, $\gamma > 1$ is assumed so that $1 < (1+\gamma)/2 < \gamma$.

Proposition 2 implies that a disfavored player is willing to engage in the mediation process even when the mediator is biased against him. In fact, the optimal mechanisms of moderately biased mediators and of the unbiased mediator coincide when $\gamma/2 < \lambda < \gamma$. So the result that all those mediators are equilibrium outcomes is obvious. However, when $\frac{\gamma(1+\gamma)}{2(2+\gamma)} \leq \lambda \leq \gamma/2$, so that high types are neither too likely nor too unlikely, the optimal mechanisms of moderately biased mediators are more complex than the optimal unbiased mechanism. Yet, those biased mediators are accepted in equilibrium; in particular, they are acceptable to the disfavored player despite the fact that such mediators may mediate against his interests.

The information rent is again the key to understanding the intuition for this unexpected result. Suppose that the given β-mediator is biased in favor of player 1 with a degree $\beta \leq \hat{\beta}$. Recall that in the optimal mechanism, the participation constraint for player 2 of type H binds; that is, he is indifferent between participating in biased mediation and engaging in war. Player 2 of type L may have an incentive to pretend to be an H-type; to prevent this, the biased mediator must provide some information rents to the L-type of disfavored player even though the mediator may wish to allocate all rents to her favored player. In particular, when $\lambda \geq \frac{\gamma(1+\gamma)}{2(2+\gamma)}$, player 2 of type L receives the information rent that makes him gain under mediation than in conflict. Thus the disfavored player would agree to accept such a mediator who is biased against him.

Celik and Peters (2011) examine a larger class of equilibria where rejection of a mechanism may be on the equilibrium path (i.e., type dependence of selection decisions). They establish the existence of mechanisms which are implementable only if the mechanisms will be rejected by some types of some players. However, in our setting, there do not exist equilibria where both selection decisions (accepting and rejecting) are on the equilibrium path. We can directly apply Proposition 2 of Celik and Peters (2011) to show that considering rejections on the equilibrium path does not extend the set of implementable outcomes in our setting.[19]

Remark 2. Using the notations of Celik and Peters (2011), we define π_i as the function that maps the type of player i and the belief system to the expected payoff in the peace-

[19]In the online appendix, we take an alternative route to show that there exists no separating equilibrium in which players use separating (type-dependent) selection strategies.

maximizing equilibrium of the default game. That is, $\pi_i(t_i, q_i, q_{-i})$ is the expected payoff for type t_i when the default game is played under belief system (q_i, q_{-i}). Let $\breve{\pi}_i(t_i, q_i, q_{-i})$ denote the biconjugate function for $\pi_i(t_i, q_i, q_{-i})$, defined as $\breve{\pi}_i(t_i, q_i, q_{-i}) = \inf\{x : (q_{-i}, x) \in co\,(\pi_i(t_i, q_i, \cdot))\}$, where $co\,(\pi_i(t_i, q_i, \cdot))$ refers to the convex hull of the graph of the function $\pi_i(t_i, q_i, \cdot)$ for fixed values of t_i and q_i. Following from Corollary 2, under the interim belief on the type of player $-i$, $q_{-i} = q$, which satisfies $q \leq (1-q)\gamma$, we have $\pi_i(L, q_i, q) = q(1-p)\theta + (1-q)(1/2)$ and $\pi_i(H, q_i, q) = q\theta/2 + (1-q)p\theta$ in the peace-maximizing equilibrium of the default game. These functions are linear in q, thus the values of functions $\breve{\pi}_i$ and π_i coincide for the interim belief $q_{-i} = q$ for all types of all players regardless of the belief q_i. Then Proposition 2 of Celik and Peters (2011) applies: any implementable allocation rule is also truthfully implementable. Truthful implementation demands for a unanimous acceptance of our mediator's optimal direct-revelation mechanism by all types of all players. Thus we only need to consider equilibria in which the given β-mediator is unanimously accepted by all types of all players.

4.2 Effectiveness of Biased Mediators

Despite the differences in the optimal mechanisms between unbiased and biased mediators, a biased mediator does not necessarily perform worse than the unbiased mediator in terms of peace probabilities.

Proposition 3. *Given the prior beliefs about the players' types,*

(1) For $\gamma > 1$ and $\lambda \leq (\gamma - 1)/2$, the optimal mechanisms of β-mediators for all $\beta \in [0, 1]$ achieve the same ex ante chance of peace.

(2) For $(\gamma - 1)/2 < \lambda < \gamma$, the optimal mechanisms of β-mediators for all $\beta \in [1 - \hat{\beta}, \hat{\beta}]$ achieve the same ex ante chance of peace, whereas the ex ante chance of peace is strictly lower under the optimal mechanisms of β-mediators with $\beta < 1 - \hat{\beta}$ or $\beta > \hat{\beta}$ than under the optimal mechanisms of β-mediators with $\beta \in [1 - \hat{\beta}, \hat{\beta}]$.

When the intensity or cost of war is low ($\gamma > 1$) and high types are very unlikely ($\lambda \leq (\gamma - 1)/2$), any biased mediator, were it unanimously selected into mediation, is exactly as effective in promoting peace as the unbiased counterpart. When war is expected to be very costly or intense ($\gamma \le 1$) or when the players' type is likely high ($\lambda > (\gamma - 1)/2$), any moderately biased mediator is exactly as effective as the unbiased counterpart. This result is particularly surprising because the unbiased mediator cares only about maximizing the ex ante peace chance, whereas a biased mediator cares relatively less about it. The crucial factor in understanding the intuition behind this result is that war is costly: war shrinks the value of the resource to $\theta < 1$ and the value of $(1 - \theta)$ is lost in the event of war. By using some portion of the resource that would have been lost, a biased mediator may offer more shares to her favored player. But the mediator's incentive to do so is limited by the need to give the information rent to an L-type of disfavored player in order to keep in check the temptation of mis-reporting by such a player. A moderately biased mediator does not find this information rent costly, thus ends up effectively achieving the same peace chance as the unbiased mediator does. However, a strongly biased mediator would find the information rent too costly, thus allocating information rents only to the favored player and yielding a lower chance of peace, unless war is not too costly ($\gamma > 1$) and high types are very unlikely.

As is demonstrated in Proposition 3, a strongly biased mediator performs (weakly) worse than any other β-mediator with $\beta \in (1-\hat{\beta}, \hat{\beta})$ in achieving peace. Mediation by such strongly biased mediators may be a concern if the goal in the dispute is to minimize the chance of war. This result seems to be in line with some academic works, both theoretical and empirical, arguing that mediator bias undercuts mediation. However, Proposition 2 has shown that acceptance of any strongly biased mediator is never an equilibrium. Thus Propositions 2 and 3 together imply the following.

Corollary 1. *In any equilibrium where a biased mediator is accepted unanimously, such a biased mediator is equally effective in minimizing the ex ante chance of war as the unbiased mediator were it given instead and accepted in equilibrium.*

We end this section with the discussion of the link between the initiation of mediation and its effectiveness. Mediator bias is related to mediator effectiveness, which in turn may affect the decision of disputants to initiate mediation in the first place. Mediators that are strongly biased towards one player are relatively less effective at resolving disputes than unbiased or moderately biased mediators, if they were unanimously selected into mediation; but mediation by strongly biased mediators is never an equilibrium outcome. One can identify an equilibrium where mediation always takes place given an unbiased mediator; and when high types are likely, mediation by moderately biased mediators can be an equilibrium outcome. Our results imply a selection effect: unbiased and moderately biased mediators that are relatively more effective at bringing about peaceful settlements are more likely to be selected into mediation, whereas strongly biased mediators that are less effective are forgone.

The intuition lies in how mediator bias affects the optimal recommendation strategies. An unbiased mediator uses the same probability of recommending peaceful settlements to asymmetric type profiles and allocates the information rents equally across the two players. Biased mediators, on the other hand, allocate more information rents to the favored player while providing a higher chance of peaceful settlements to the disfavored player; and strongly biased mediators allocate no information rent to the disfavored player. When making a decision at the selection stage, a player would expect such mediation outcomes if given a strongly biased mediator who is biased against him. In particular, when he is of L-type, he would anticipate that the equilibrium outcome under mediation will be no different than the outcome without mediation. Because he might enjoy peace with some positive probability in the default game, he would rather refuse mediation by any strongly biased mediator.

Our discussion here, however, is not to say that there is a causal effect of mediator bias (thus its effectiveness) on mediator selection. In fact, any (biased or unbiased) mediator's optimal mechanism cannot be worse than any equilibrium without the mediator.[20] Even the strongly biased mediators, although they fail to be selected in equilibrium due to rejection by one party, achieve a higher ex ante chance of peace than what the players would achieve in the default game. Further, a unanimous acceptance of any moderately biased mediator is not an equilibrium when the H/L odds ratio is relatively low ($\lambda < \frac{\gamma(1+\gamma)}{2(2+\gamma)}$), even though such a mediator, were it selected, would be equally effective as the unbiased counterpart.

[20] A mediator could always, trivially, make the messages she receives public, thereby mimicking the optimal unmediated- or no-communication equilibrium.

5 Concluding Remarks

In this paper, we provide novel results that link mediator bias to both the selection and the effectiveness of mediation. We show that, compared to the unbiased mediator, biased mediators use recommendation strategies that induce more interim payoffs to the favored player while guaranteeing more chances of peace to the disfavored player. When the chance of facing a high-type player is relatively high, the players accepting *biased* mediation with probability one can be an equilibrium. When the players' type is *more* likely high, a player would be willing to accept mediation by a more strongly biased mediator. Further, we surprisingly find that the biased mediator, if unanimously accepted, achieves the same ex ante probability of peace as the unbiased mediator. This implies that once a mediator is selected into mediation, mediator bias is independent of her effectiveness in resolving conflicts. A biased mediator cares more about the payoff of her favored player than about maximizing the ex ante chance of peace. So the biased mediator would want to allocate more resource to her favored player but at the same time she needs to elicit the players' private information truthfully during the mediation process. This increases the need to provide information rents to the disfavored player, and the biased mediator does so by recommending peaceful settlement to the low-type disfavored player more often than would the unbiased mediator do. Thus the biased mediator ends up being exactly as effective in promoting peace as the unbiased mediator.

Appendix A Default Game under Arbitrary Beliefs

In this appendix, we analyze the default game described in Section 2.2 under an arbitrary belief system represented by $\{q_1, q_2\}$. For each $i = 1, 2$, q_i yields the probability that player $-i$ assigns to the event that player i has type H when the players are playing the default game. The choice variable for player i is its announcement of peace or war, i.e., $a_i \in \{P, W\}$, which can be conditioned on the realization of its type. Here we will use $\alpha_i^{t_i}$ to denote the probability that player i of type $t_i \in \{H, L\}$ announces P.

The expected payoff of player i of type t_i under the belief q_{-i} given a strategy profile $\alpha = \{(\alpha_i^{t_i})_{t_i \in T_i}\}_{i\in\{1,2\}}$, denoted by $u_i(\alpha|t_i)$, can be computed as follows:

$$\begin{aligned} u_i(\alpha|H) =& q_{-i}\left(\alpha_i^H\alpha_{-i}^H(1/2) + (1-\alpha_i^H\alpha_{-i}^H)(\theta/2)\right) + (1-q_{-i})\left(\alpha_i^H\alpha_{-i}^L(1/2) + (1-\alpha_i^H\alpha_{-i}^L)p\theta\right);\\ u_i(\alpha|L) =& q_{-i}\left(\alpha_i^L\alpha_{-i}^H(1/2) + (1-\alpha_i^L\alpha_{-i}^H)(1-p)\theta\right)\\ &+ (1-q_{-i})\left(\alpha_i^L\alpha_{-i}^L(1/2) + (1-\alpha_i^L\alpha_{-i}^L)(\theta/2)\right). \end{aligned}$$

To determine best responses, we compute how the payoffs change with respect to $\alpha_i^{t_i}$: $\frac{\partial u_i(\alpha|H)}{\partial \alpha_i^H} = q_{-i}\alpha_{-i}^H(1-\theta)/2 - (1-q_{-i})\alpha_{-i}^L(p\theta - 1/2) = \left[q_{-i}\alpha_{-i}^H - (1-q_{-i})\gamma\alpha_{-i}^L\right](1-\theta)/2$ and $\frac{\partial u_i(\alpha|L)}{\partial \alpha_i^L} = q_{-i}\alpha_{-i}^H(1/2 - (1-p)\theta) + (1-q_{-i})\alpha_{-i}^L(1-\theta)/2 = \left[q_{-i}\alpha_{-i}^H(\gamma+2) + (1-q_{-i})\alpha_{-i}^L\right](1-\theta)/2$, where $\gamma = \frac{p\theta - 1/2}{1/2-\theta/2}$. Because $\frac{\partial u_i(\alpha|L)}{\partial \alpha_i^L} \geq 0$ regardless of α_{-i}^H and α_{-i}^L, announcing peace with probability one is a weakly dominant strategy for an L-type, i.e., $\alpha_i^L = 1$ for $i = 1, 2$; and unless $\alpha_{-i}^H = \alpha_{-i}^L = 0$, announcing peace with probability one is a strictly dominant strategy for an L-type. The following lemma describes an H-type player's best response given that an L-type player uses his weakly dominant strategy.

Lemma 2. *Suppose that an L-type of player $-i$ uses his weakly dominant strategy, i.e., $\alpha_{-i}^L = 1$ for $i = 1, 2$. If $q_{-i} \leq (1-q_{-i})\gamma$, then announcing war with probability one is a weakly dominant strategy for player i of type H. If $q_{-i} > (1-q_{-i})\gamma$, then announcing war with probability one is the best response for player i of type H against $\alpha_{-i}^H \leq \frac{1-q_{-i}}{q_{-i}}\gamma < 1$.*

Proof. Given $\alpha_{-i}^L = 1$, we have $\frac{\partial u_i(\alpha|H)}{\partial \alpha_i^H} = \left[q_{-i}\alpha_{-i}^H - (1-q_{-i})\gamma\right](1-\theta)/2$, which is less than or equal to zero iff $q_{-i}\alpha_{-i}^H \leq (1-q_{-i})\gamma$. If $q_{-i} \leq (1-q_{-i})\gamma$, then because $\alpha_{-i}^H \leq 1$, $\alpha_i^H = 0$ is a weakly dominant strategy for player i of type H. (Note that if $\alpha_{-i}^H < 1$, then $\alpha_1^H = 0$ is a strictly dominant strategy.) If $q_{-i} > (1-q_{-i})\gamma$, then the best response for player i of type H is $\alpha_i^H = 0$ against $\alpha_{-i}^H \geq 0$ such that $q_{-i}\alpha_{-i}^H \leq (1-q_{-i})\gamma$, and $\alpha_i^H = 1$ against $\alpha_{-i}^H = 1$. $\square$

Note that there always exists an equilibrium of the default game where both types of both players always announce war regardless of their beliefs. Excluding such a trivial equilibrium, we characterize the equilibria of our default game under arbitrary beliefs. The proof follows directly from Lemma 2.

Proposition 4. *The set of Bayesian Nash equilibria of the default game under the belief system $\{q_1, q_2\}$ consists of the following:*

(1) For $q_1 \le (1-q_1)\gamma$ or $q_2 \le (1-q_2)\gamma$, L-types choose peace and H-types choose war. The equilibrium probability of peace is $(1-q_1)(1-q_2)$.

(2) For $q_1 > (1-q_1)\gamma$ and $q_2 > (1-q_2)\gamma$,

(a) L-types choose peace and H-types choose war. The equilibrium probability of peace is $(1-q_1)(1-q_2)$.

(b) Both types choose peace. The equilibrium probability of peace is one.

(c) L-types choose peace, and H-types choose peace with probability $\gamma(1-q_i)/q_i$ for $i = 1, 2$. The equilibrium probability of peace is $(1-q_1+\gamma(1-q_1))\,(1-q_2+\gamma(1-q_2))$.

We directly obtain the following corollary.

Corollary 2. *In the peace-maximizing equilibrium:*

(1) For $q_1 \le (1-q_1)\gamma$ or $q_2 \le (1-q_2)\gamma$, the expected payoff is $q_{-i}(1-p)\theta + (1-q_{-i})(1/2)$ for an L-type player i and $q_{-i}(\theta/2) + (1-q_{-i})p\theta$ for an H-type player i.

(2) For $q_1 > (1-q_1)\gamma$ and $q_2 > (1-q_2)\gamma$, the expected payoff for both types of player i is $1/2$.

If the default game is played under the interim beliefs $q_1 = q_2 = q$, then Case (1) of Proposition 4 and Corollary 2 applies.

Appendix B The Optimal Mediation Programme

In the mediation game, the β-mediator given beliefs $\{q_1, q_2\}$ is represented by her direct-revelation mechanism $\mathcal{M}(\beta, q_1, q_2) = (p_t, b_t)_{t\in T}$. Recall that b_t is the share of the resource allocated to player 1 when the reported type profile is t. Let d_t denote the share of the (remaining) resource allocated to player 1 under war; so $\theta - d_t$ is the share allocated to player 2 under war. Note that $d_{HH} = d_{LL} = \theta/2$, $d_{HL} = p\theta$, and $d_{LH} = (1-p)\theta$.

The β-mediator's payoff given the (truthfully) reported types t is characterized and rearranged as:

$$\begin{aligned}\omega_\beta(t) &= \beta\left[p_t b_t + (1-p_t)d_t\right] + (1-\beta)\left[p_t(1-b_t) + (1-p_t)(\theta - d_t)\right] \\ &= (2\beta-1)p_t(b_t - d_t) + (1-\beta)(1-\theta)p_t + \big[\beta d_t + (1-\beta)(\theta - d_t)\big].\end{aligned}$$

We denote the terms in brackets, $\big[\;\big]$, by D_t. Then the β-mediator's expected payoff given beliefs $\mathbf{q} = (q_1, q_2)$ can be written in terms of player 1's payoffs as follows.

$$\begin{aligned}W_\beta(\mathcal{M}(\beta, q_1, q_2)) = E_{\mathbf{q}}[\omega_\beta(t)] &= (2\beta-1)E_{\mathbf{q}}\left[p_t(b_t - d_t)\right] + (1-\beta)(1-\theta)E_{\mathbf{q}}[p_t] + E_{\mathbf{q}}[D_t] \\ &= (2\beta-1)\Big[q_1 q_2 p_{HH}\left(b_{HH} - \theta/2\right) + q_1(1-q_2)p_{HL}(b_{HL} - p\theta) \\ &\quad + (1-q_1)q_2 p_{LH}\left(b_{LH} - (1-p)\theta\right) + (1-q_1)(1-q_2)p_{LL}\left(b_{LL} - \theta/2\right)\Big] \\ &\quad + (1-\beta)(1-\theta)\Big[q_1 q_2 p_{HH} + q_1(1-q_2)p_{HL} \\ &\quad + (1-q_1)q_2 p_{LH} + (1-q_1)(1-q_2)p_{LL}\Big] + E_{\mathbf{q}}[D_t],\end{aligned}$$

where $E_{\mathbf{q}}[D_t] = q_1 q_2 D_{HH} + q_1(1-q_2)D_{HL} + (1-q_1)q_2 D_{LH} + (1-q_1)(1-q_2)D_{LL}$. We may also equivalently express $W_\beta(\mathcal{M}(\beta, q_1, q_2))$ in terms of player 2's payoffs.

Three cases immediately follow from the observation of the β-mediator's expected payoff. If $\beta = 1$, then $W_{\beta=1} = E_{\mathbf{q}}[p_t b_t + (1-p_t)d_t]$, which is equal to player 1's ex ante expected payoff. If $\beta = 0$, then $W_{\beta=0} = E_{\mathbf{q}}[p_t(1-b_t) + (1-p_t)(\theta - d_t)]$, which is equal to player 2's ex ante expected payoff. If $\beta = 1/2$, then $W_{\beta=1/2} = (1/2)(1-\theta)E_{\mathbf{q}}[p_t] + E_{\mathbf{q}}[D_t] = (1/2)(1-\theta)E_{\mathbf{q}}[p_t] + \theta/2$ because $D_t = \theta/2$ for all t. Note that $E_{\mathbf{q}}[p_t]$ is the ex ante probability of peace under $\mathcal{M}(1/2, \mathbf{q})$. Thus $W_{\beta=1/2}$ is equal to the ex ante probability of peace up to a multiplicative constant $(1/2)(1-\theta)$ and an additive constant $\theta/2$.

The term $E_{\mathbf{q}}[D_t]$ is a constant given β, which does not affect the mediator's optimization programme and can be dropped. Let B_t denote the ratio of player 1's gain or loss from the mediator's recommended split relative to war over the lost resource from war for a given type profile $t \in T$: $B_t = \frac{b_t - d_t}{1-\theta}$. The B_t represents player 1's net gain or loss from mediation instead of waging war (conditional on participation and truthful revelation in mediation).

In Section 3.2, we focus on the case of unanimous acceptance by all types of all players. The posterior beliefs about the players' types are identical to the interim beliefs, $q_1 = q_2 = q$. Recall that $\lambda = \frac{q}{1-q}$ is the H/L-type odds ratio and that $\lambda < \gamma$. The optimal mediation programme **(P)** of β-mediator given the belief λ (about both players) can then be reformulated as follows:

$$\textbf{(P)} \quad \max_{(p_t,B_t)_{t\in T}} \tilde{W}_\beta \left((p_t, B_t)_{t\in T}\right) = (2\beta-1)\Big[\lambda\lambda p_{HH}B_{HH} + \lambda p_{HL}B_{HL} + \lambda p_{LH}B_{LH} + p_{LL}B_{LL} \\ + (1-\beta)\Big[\lambda\lambda p_{HH} + \lambda p_{HL} + \lambda p_{LH} + p_{LL}\Big]$$

subject to

$$\begin{aligned}
(IC_{1H}) &\quad \lambda p_{HH}B_{HH} + p_{HL}B_{HL} \geq \lambda p_{LH}\left(B_{LH} - (1+\gamma)/2\right) + p_{LL}\left(B_{LL} - (1+\gamma)/2\right), \\
(IC_{1L}) &\quad \lambda p_{LH}B_{LH} + p_{LL}B_{LL} \geq \lambda p_{HH}\left(B_{HH} + (1+\gamma)/2\right) + p_{HL}\left(B_{HL} + (1+\gamma)/2\right), \\
(IC_{2H}) &\quad \lambda p_{HH}\left(1 - B_{HH}\right) + p_{LH}(1 - B_{LH}) \\
&\quad \geq \lambda p_{HL}\left(1 - B_{HL} - (1+\gamma)/2\right) + p_{LL}\left(1 - B_{LL} - (1+\gamma)/2\right), \\
(IC_{2L}) &\quad \lambda p_{HL}(1 - B_{HL}) + p_{LL}\left(1 - B_{LL}\right) \\
&\quad \geq \lambda p_{HH}\left(1 - B_{HH} + (1+\gamma)/2\right) + p_{LH}\left(1 - B_{LH} + (1+\gamma)/2\right), \\
(PC_{1H}) &\quad \lambda p_{HH}B_{HH} + p_{HL}B_{HL} \geq 0, \\
(PC_{1L}) &\quad \lambda p_{LH}B_{LH} + p_{LL}B_{LL} \geq 0, \\
(PC_{2H}) &\quad \lambda p_{HH}\left(1 - B_{HH}\right) + p_{LH}(1 - B_{LH}) \geq 0, \\
(PC_{2L}) &\quad \lambda p_{HL}(1 - B_{HL}) + p_{LL}\left(1 - B_{LL}\right) \geq 0, \\
\text{and to} &\quad 0 \leq p_t \leq 1 \text{ and } -d_t/(1-\theta) \leq B_t \leq (1-d_t)/(1-\theta) \text{ for all } t \in T.
\end{aligned}$$

Appendix C Proof of Proposition 1

If all types of all players choose to accept mediation, then the posterior beliefs on the players' types are $q_1 = q_2 = q$. Because $q \in (0,1)$, we use the parameter $\lambda = \frac{q}{1-q} < \gamma$, as is assumed in the main text. We proceed by proving two lemmas and analyzing different cases in parts.

Lemma 3. *The participation constraints for L-types, (PC_{iL}), $i = 1, 2$, are trivially satisfied* *if the participation constraints for H-types, (PC_{iH}), $i = 1, 2$, are satisfied.*

Proof. We write $LHS(\cdot)$ for the left-hand-side of a constraint and $RHS(\cdot)$ for the right-hand-side of a constraint. By observing the constraints (IC_{iL}), (IC_{iH}), and (PC_{iH}), we can see that for each player $i = 1, 2$, $LHS(PC_{iL}) \geq RHS(IC_{iL}) \geq LHS(IC_{iH}) = LHS(PC_{iH}) \geq 0$. $\square$

We shall henceforth state the results for the case of $\beta > 1/2$; the results for the case $\beta < 1/2$ can be analogously stated by interchanging the roles of players 1 and 2.

We first formulate the following relaxed programme without the constraint (IC_{2H}) and the two non-violated constraints (PC_{1L}) and (PC_{2L}):

$$\max_{(p_t,B_t)_{t\in T}} \tilde{W}_\beta \left((p_t, B_t)_{t\in T}\right) = (2\beta - 1)\Big[\lambda\lambda p_{HH}B_{HH} + \lambda p_{HL}B_{HL} + \lambda p_{LH}B_{LH} + p_{LL}B_{LL}$$
$$+ (1-\beta)\Big[\lambda\lambda p_{HH} + \lambda p_{HL} + \lambda p_{LH} + p_{LL}\Big]$$

subject to

$$(IC_{1H}) \quad \lambda p_{HH}B_{HH} + p_{HL}B_{HL} \geq \lambda p_{LH}\left(B_{LH} - (1+\gamma)/2\right) + p_{LL}\left(B_{LL} - (1+\gamma)/2\right),$$

$$(IC_{1L}) \quad \lambda p_{LH}B_{LH} + p_{LL}B_{LL} \geq \lambda p_{HH}\left(B_{HH} + (1+\gamma)/2\right) + p_{HL}\left(B_{HL} + (1+\gamma)/2\right),$$

$$(IC_{2L}) \quad \lambda p_{HL}(1 - B_{HL}) + p_{LL}\left(1 - B_{LL}\right)$$
$$\geq \lambda p_{HH}\left(1 - B_{HH} + (1+\gamma)/2\right) + p_{LH}\left(1 - B_{LH} + (1+\gamma)/2\right),$$

$$(PC_{1H}) \quad \lambda p_{HH}B_{HH} + p_{HL}B_{HL} \geq 0,$$

$$(PC_{2H}) \quad \lambda p_{HH}\left(1 - B_{HH}\right) + p_{LH}(1 - B_{LH}) \geq 0,$$

and to the relaxed constraints $p_t \geq 0$ and $B_t \geq -d_t/(1-\theta)$ for all $t \in T$.

Lemma 4. *The participation constraint for an H-type of player 2, (PC_{2H}), and the incentive compatibility constraints for L-types of both players, (IC_{iL}), $i = 1, 2$, bind in the solution of the above relaxed programme.*

Proof. First, suppose that (PC_{2H}) does not bind. Then it is possible to increase both B_{HH} and B_{LH} thus increasing the value of the objective function $\tilde{W}_\beta$ in a way that does not violate (PC_{2H}) and keeps in check the other constraints; in particular, increasing B_{HH} and B_{LH} by the amount that leaves the constraints (IC_{1H}) and (IC_{1L}) intact makes (IC_{2L}) and

(PC_{1H}) slack. Second, (IC_{2L}) must bind in the solution, or else one could increase both B_{HL} and B_{LL}, thus increasing the objective function, in a way that does not violate (IC_{2L}), increases each side of the constraints (IC_{1H}) and (IC_{1L}) by the same amount, and makes (PC_{1H}) slack. Finally, (IC_{1L}) must bind as well. Suppose it does not. Note that (PC_{1H}) implies that at least one of B_{HH} or B_{HL} is non-negative. There are three cases to consider. Case 1 ($B_{HL} < 0$): One could increase both p_{HL} and B_{HL} without violating (IC_{1L}) in a way that leaves the left hand side of (IC_{2L}) unchanged, makes (IC_{1H}) and (PC_{1H}) slack, and increases the objective function. Case 2 ($B_{HL} \in [0, 1]$): Then it is possible to increase p_{HL}, thus increasing the objective function without violating (IC_{1L}), which makes (IC_{1H}), (IC_{2L}), and (PC_{1H}) slack. Case 3 ($B_{HL} > 1$): It must be that $B_{LL} < 1$ because Lemma 3 implies for the non-binding (PC_{2L}) that at least one of $(1 - B_{HL})$ or $(1 - B_{LL})$ is strictly positive. If $B_{LL} \leq 0$, then one could increase p_{LL} without violating (IC_{1L}), thus increasing the objective function and making (IC_{1H}) and (IC_{2L}) slack. If $B_{LL} \in (0, 1)$, then one could increase both p_{HL} and p_{LL} without violating the other constraints. $\square$

Taking into account the binding constraints and after rearranging, we can rewrite the relaxed programme as follows, which we will refer to as **(RP)**:

$$\textbf{(RP)} \quad \max_{(p_t, B_t)_{t\in T}} W_\beta((p_t)_{t\in T}) \equiv \beta\left(\lambda p_{HL} + p_{LL}\right) + \left[\beta\lambda - (2\beta - 1)(1+\gamma)/2\right](\lambda p_{HH} + p_{LH})$$

subject to the relaxed constraints $p_t \geq 0$ and $B_t \geq -d_t/(1-\theta)$ for all $t \in T$;

$$(IC_{1H}) \quad \lambda p_{LH} + p_{LL} \geq \lambda p_{HH} + p_{HL},$$

$$(PC_{1H}) \quad p_{LL} - \left((1+\gamma)/2 - \lambda\right) p_{LH} \geq \lambda(1+\gamma-\lambda)p_{HH} + \left((1+\gamma)/2 - \lambda\right) p_{HL};$$

and to the binding constraints determining B_t for $t \in T$:

$$(IC_{1L}) \quad \lambda p_{LH} B_{LH} + p_{LL} B_{LL} = \lambda p_{HH}\left(B_{HH} + (1+\gamma)/2\right) + p_{HL}\left(B_{HL} + (1+\gamma)/2\right),$$

$$(IC_{2L}) \quad \lambda p_{HL}(1 - B_{HL}) + p_{LL}\left(1 - B_{LL}\right) = (\lambda p_{HH} + p_{LH})(1+\gamma)/2,$$

$$(PC_{2H}) \quad \lambda p_{HH}\left(1 - B_{HH}\right) + p_{LH}(1 - B_{LH}) = 0.$$

Taking into account the binding constraints (IC_{2L}) and (PC_{2H}), the missing constraint

(IC_{2H}) in the relaxed programme can be rewritten as

$$(IC_{2H}) \quad \lambda p_{HL} + p_{LL} \geq \lambda p_{HH} + p_{LH}.$$

We note that $p_{LL} = 1$ in the solution because setting $p_{LL} = 1$ maximizes the left-hand-sides of (IC_{1H}) and (PC_{1H}), and does not affect their right-hand-sides.

We proceed in two steps: In Step 1, we take p_{HH} and p_{LH} as fixed and solve for the maximal value of $p_{HL} \in [0,1]$ that satisfies the two constraints (IC_{1H}) and (PC_{1H}), formulated as **(RP1)** below. In Step 2, for each possible maximal p_{HL}, we characterize the feasible values of $p_{HH} \in [0,1]$ and $p_{LH} \in [0,1]$ that satisfy the relevant constraints, and choose the optimal values of p_{HH} and p_{LH} conditional on the given p_{HL}; we then determine the optimal solution of the relaxed programme **(RP)**. Finally, we check that such a solution satisfies all the constraints of the unrelaxed optimal mediation programme **(P)**.

Step 1. For fixed values of $p_{HH} \in [0,1]$ and $p_{LH} \in [0,1]$, consider the following problem:

$$\begin{aligned}
\textbf{(RP1)} \quad & \max_{p_{HL}\in[0,1]} \quad \beta(\lambda p_{HL} + 1) \\
\text{s.t.} \quad & (IC_{1H}) \quad p_{HL} \leq 1 - \lambda(p_{HH} - p_{LH}), \\
& (PC_{1H}) \quad ((1+\gamma)/2 - \lambda)\, p_{HL} \leq 1 - \lambda(1+\gamma-\lambda)p_{HH} - ((1+\gamma)/2 - \lambda)\, p_{LH}.
\end{aligned}$$

If $\lambda < (1+\gamma)/2$, then the feasible set for p_{HL} is characterized by

$$0 \leq p_{HL} \leq \min\left\{1, 1 - \lambda(p_{HH} - p_{LH}), \frac{1 - \lambda(1+\gamma-\lambda)p_{HH} - ((1+\gamma)/2 - \lambda)p_{LH}}{(1+\gamma)/2 - \lambda}\right\}.$$

If $\lambda = (1+\gamma)/2$, then the feasible set for p_{HL} is characterized by

$$0 \leq p_{HL} \leq \min\left\{1, 1 - \lambda(p_{HH} - p_{LH})\right\}.$$

If $(1+\gamma)/2 < \lambda < \gamma$, then the feasible set for p_{HL} is characterized by

$$\max\left\{0, \frac{1 - \lambda(1+\gamma-\lambda)p_{HH} - ((1+\gamma)/2 - \lambda)p_{LH}}{(1+\gamma)/2 - \lambda}\right\} \leq p_{HL} \leq \min\left\{1, 1 - \lambda(p_{HH} - p_{LH})\right\}.$$

Because the objective function increases in p_{HL}, the optimal solution for p_{HL} in the relaxed problem **(RP1)**, if it exists, occurs at the supremum of the feasible set:

$$p_{HL} = \min\Big\{1, 1-\lambda(p_{HH}-p_{LH}), \frac{1-\lambda(1+\gamma-\lambda)p_{HH}-((1+\gamma)/2-\lambda)p_{LH}}{(1+\gamma)/2-\lambda}\Big\}, \quad \text{(C.1)}$$

where the last component is present only if $\lambda < (1+\gamma)/2$.

Remark 3. For each possible maximal value of p_{HL}, given $p_{LL} = 1$, the objective function $W_\beta((p_t)_{t\in T})$ in **(RP)** can be rewritten as follow. Recall that we are focusing on $\beta > 1/2$.

(1) $p_{HL} = 1$:

$$W_\beta^{(1)} \equiv \beta(\lambda+1) + \big[(1+\gamma)/2 - \beta(1+\gamma-\lambda)\big](\lambda p_{HH} + p_{LH}).$$

If $\beta < \hat{\beta} \equiv \frac{1+\gamma}{2(1+\gamma-\lambda)}$, then the coefficients on p_{HH} and p_{LH} are strictly positive.

(2) $p_{HL} = 1-\lambda(p_{HH}-p_{LH})$:

$$W_\beta^{(2)} \equiv \beta(\lambda+1) - (2\beta-1)((1+\gamma)/2)\lambda p_{HH} + \big[(1+\gamma)/2 - \beta(1+\gamma-\lambda-\lambda^2)\big]p_{LH}.$$

The coefficient on p_{HH} is strictly negative. If $\beta \le \hat{\beta}$, then the coefficient on p_{LH} is strictly positive.

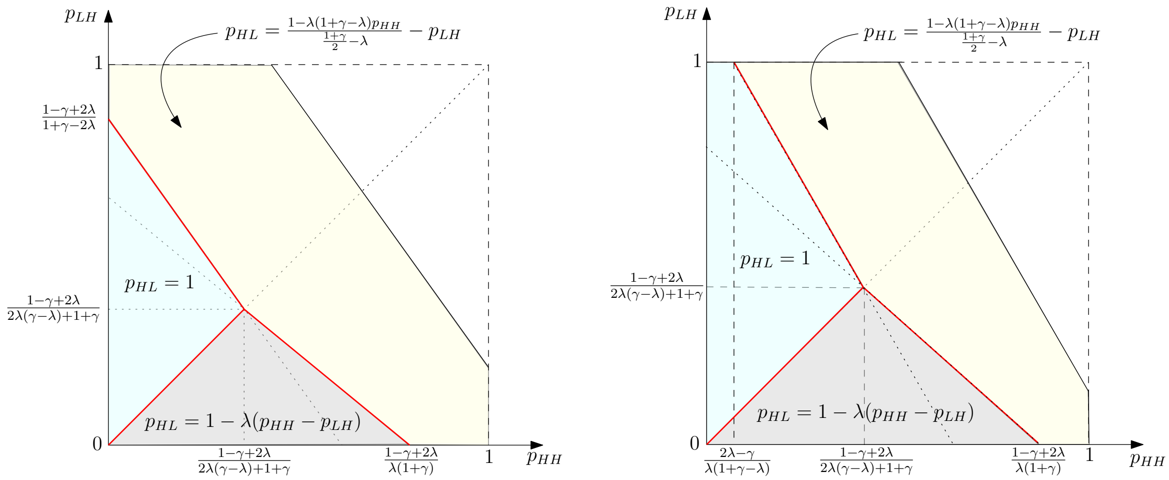
(3) $p_{HL} = \frac{1-\lambda(1+\gamma-\lambda)p_{HH}-((1+\gamma)/2-\lambda)p_{LH}}{(1+\gamma)/2-\lambda}$ (which can be possible only if $\lambda < (1+\gamma)/2$):

$$W_\beta^{(3)} \equiv \big((1+\gamma)/2\big)\Big[\frac{\beta}{(1+\gamma)/2-\lambda} + \lambda\Big(1-\frac{\beta(1+\gamma-\lambda)}{(1+\gamma)/2-\lambda}\Big)p_{HH} + (1-2\beta)p_{LH}\Big].$$

The coefficients on p_{HH} and p_{LH} are both strictly negative, because $(1+\gamma)/2-\lambda-\beta(1+\gamma-\lambda) = (1-2\beta)(1+\gamma)/2 - \lambda(1-\beta) < 0$ and $1-2\beta < 0$.

Step 2. We consider two cases: **(A)** $\lambda \le (1+\gamma)/2$ and **(B)** $(1+\gamma)/2 < \lambda < \gamma$.

Case **(A)** $\lambda \le (1+\gamma)/2$: For each possible p_{HL} characterized in (C.1), the feasible set for (p_{HH}, p_{LH}) is illustrated in Figure 2.



(a) $\lambda \leq \frac{\gamma}{2}$ (b) $\lambda > \frac{\gamma}{2}$

Figure 2: Feasible set for (p_{HH}, p_{LH}) when $\lambda \leq (1+\gamma)/2$. *Note*: In the figure, $(1-\gamma+2\lambda) > 0$ and $\gamma > 1$ (so that $\frac{1-\gamma+2\lambda}{\lambda(1+\gamma)} < 1$) are assumed.

(1) If $p_{HL} = 1$, then the feasible set for (p_{HH}, p_{LH}) is defined by the following inequalities from (IC_{1H}) and (PC_{1H}):

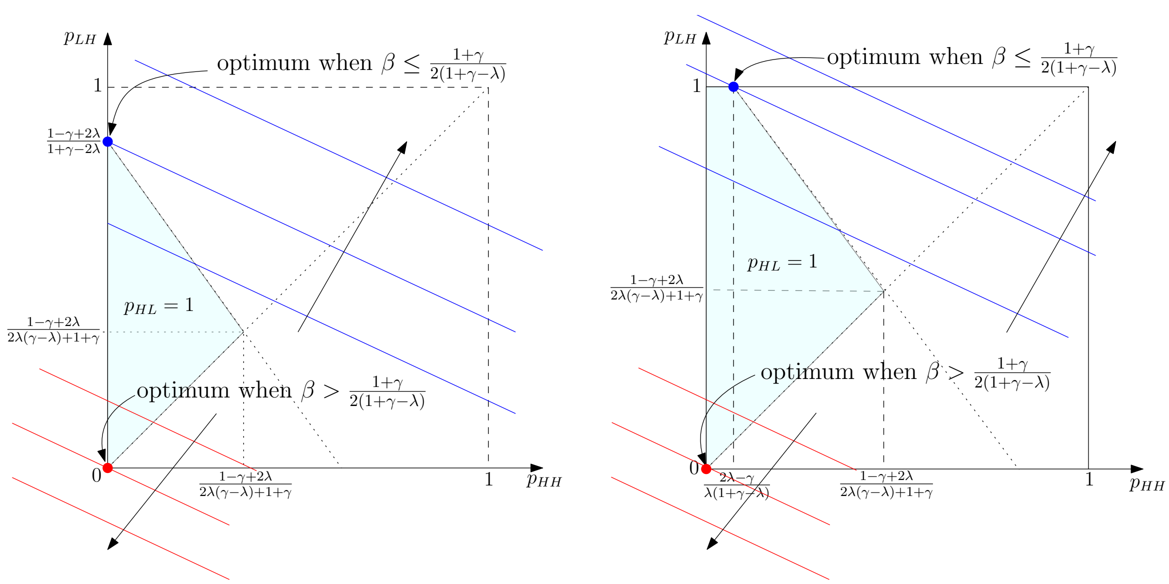
$$\begin{aligned} & p_{LH} \geq p_{HH}; \\ & p_{LH} \leq -\frac{2\lambda(1+\gamma-\lambda)}{1+\gamma-2\lambda} p_{HH} + \frac{1-\gamma+2\lambda}{1+\gamma-2\lambda} \quad \text{if } \lambda < (1+\gamma)/2, \\ & p_{HH} \leq \frac{1}{\lambda(1+\gamma-\lambda)} \quad \text{if } \lambda = (1+\gamma)/2. \end{aligned}$$

(i) For $\beta < \hat{\beta} \equiv \frac{1+\gamma}{2(1+\gamma-\lambda)}$, both p_{HH} and p_{LH} increase the objective function $W_\beta^{(1)}$, but they cannot both be one due to the constraint (PC_{1H}). Because the marginal rate of substitution (MRS) of p_{LH} for p_{HH}, represented by λ, is less than the absolute value of the coefficient on p_{HH} in (PC_{1H}), $\frac{2\lambda(1+\gamma-\lambda)}{1+\gamma-2\lambda}$ (given $\lambda < (1+\gamma)/2$), increasing p_{LH} always contributes more to the objective function than increasing p_{HH} does. Thus the optimum occurs at the point that maximizes p_{LH}. When $\lambda = (1+\gamma)/2$, setting $p_{LH} = 1$ and $p_{HH} = \min\{1, \frac{1}{\lambda(1+\gamma-\lambda)}\}$ (from the binding (PC_{1H}) and $p_{HH} \leq 1$) maximizes the objective function. When $\gamma/2 < \lambda < (1+\gamma)/2$, setting $p_{LH} = 1$ in (PC_{1H}) implies $p_{HH} \leq \frac{2\lambda-\gamma}{\lambda(1+\gamma-\lambda)}$; hence, $p_{LH} = 1$ and $p_{HH} = \frac{2\lambda-\gamma}{\lambda(1+\gamma-\lambda)} \in (0,1)$ from the binding (PC_{1H})

constraint. When $\lambda \leq \gamma/2$, setting $p_{LH} = 1$ makes p_{HH} negative. Hence, we set $p_{HH} = 0$ and obtain $p_{LH} = \frac{1-\gamma+2\lambda}{1+\gamma-2\lambda} \leq 1$ from the binding (PC_{1H}) constraint. We can verify that the quantity p_{LH} is strictly positive conditional on the optimal value of p_{HL} being one (which essentially can occur only when $1-\gamma+2\lambda$). That is, if $1-\gamma+2\lambda \leq 0$, we must set $p_{LH} = 0$ in which case $1 \geq \frac{1}{(1+\gamma)/2-\lambda}$, which contradicts $p_{HL} = 1$ in (C.1).

(ii) For $\beta \geq \hat{\beta}$, both p_{HH} and p_{LH} decrease the objective function $W_\beta^{(1)}$. Hence, $p_{HH} = p_{LH} = 0$ at the optimum.

Figure 3 illustrates the optimal values of (p_{HH}, p_{LH}) conditional on $p_{HL} = 1$.



(a) $\lambda \leq \frac{\gamma}{2}$ (b) $\lambda > \frac{\gamma}{2}$

Figure 3: The optimal values of (p_{HH}, p_{LH}) when $\lambda \leq (1+\gamma)/2$ given $p_{HL} = 1$

(2) If $p_{HL} = 1 - \lambda(p_{HH} - p_{LH})$, then the (IC_{1H}) constraint binds and the feasible set for (p_{HH}, p_{LH}) is characterized by:

$$
\begin{aligned}
&p_{LH} \leq p_{HH};\\
&p_{LH}(\lambda+1) \leq -\frac{\lambda(1+\gamma)}{1+\gamma-2\lambda}p_{HH} + \frac{1-\gamma+2\lambda}{1+\gamma-2\lambda}, \quad \text{if } \lambda < (1+\gamma)/2,\\
&p_{HH} \leq \frac{1}{\lambda(1+\gamma-\lambda)} \quad \text{if } \lambda = (1+\gamma)/2,
\end{aligned}
$$

where the first inequality follows from $p_{HL} \leq 1$ and the binding (IC_{1L}), and the second and third inequalities are (PC_{1H}).

(i) For $\beta \leq \hat{\beta}$, in the objective function $W_\beta^{(2)}$, the coefficient on p_{HH} is strictly negative and the coefficient on p_{LH} is strictly positive. So maximization of the objective function requires maximization of p_{LH} and minimization of p_{HH}. When $\lambda = (1+\gamma)/2$, we want to set p_{LH} high as possible and p_{HL} low as possible such that $p_{LH} \leq p_{HH} \leq \frac{1}{\lambda(1+\gamma-\lambda)}$; in the objective function, the coefficient on p_{LH} is larger that the absolute value of the coefficient on p_{HH}, that is, $\left[(1+\gamma)/2 - \beta(1+\gamma-\lambda-\lambda^2)\right] > (2\beta-1)((1+\gamma)/2)\lambda$, and so increasing p_{LH} contributes more to the objective function than reducing p_{HH}. Thus, the optimum occurs at $p_{LH} = p_{HH} = \min\{1, \frac{1}{\lambda(1+\gamma-\lambda)}\} \in (0,1]$. When $\lambda < (1+\gamma)/2$, the optimum occurs at $p_{LH} = p_{HH}$ and the binding (PC_{1H}), which gives $p_{LH} = p_{HH} = \frac{1-\gamma+2\lambda}{2\lambda(\gamma-\lambda)+1+\gamma} \in (0,1)$. Note that at the optimum, $p_{HL} = 1$ and both (IC_{1H}) and (PC_{1H}) bind. If $1-\gamma+2\lambda \leq 0$, then the optimum occurs at $p_{LH} = p_{HH} = 0$ in which case $p_{HL} = 1 \geq \frac{1}{(1+\gamma)/2-\lambda}$, contradicting (C.1).

(ii) For $\beta > \hat{\beta}$, the coefficient on p_{HH} strictly negative, and the coefficient on p_{LH} may be positive or negative. If both coefficients are negative, then $p_{HH} = p_{LH} = 0$ at the optimum. Even when the coefficient on p_{LH} is positive, because the MRS of p_{HH} for p_{LH} is greater than one, setting $p_{HH} = p_{LH} = 0$ maximizes the objective function. At the optimum, $p_{HL} = 1$, and (IC_{1H}) binds while (PC_{1H}) does not.

Figure 4 illustrates the optimal values of (p_{HH}, p_{LH}) conditional on $p_{HL} = 1-\lambda(p_{HH} = p_{LH})$.

(3) If $p_{HL} = \frac{1-\lambda(1+\gamma-\lambda)p_{HH}-((1+\gamma)/2-\lambda)p_{LH}}{(1+\gamma)/2-\lambda}$, which is possible only when $\lambda < (1+\gamma)/2$, then (PC_{1H}) binds, and the feasible set for (p_{HH}, p_{LH}) is characterized by:

$$\begin{aligned}
p_{LH} &\leq -\frac{2\lambda(1+\gamma-\lambda)}{1+\gamma-2\lambda}p_{HH} + \frac{2}{1+\gamma-2\lambda}, \\
p_{LH} &\geq -\frac{2\lambda(1+\gamma-\lambda)}{1+\gamma-2\lambda}p_{HH} + \frac{1-\gamma+2\lambda}{1+\gamma-2\lambda}, \\
p_{LH}(\lambda+1) &\geq -\frac{\lambda(1+\gamma)}{1+\gamma-2\lambda}p_{HH} + \frac{1-\gamma+2\lambda}{1+\gamma-2\lambda},
\end{aligned}$$

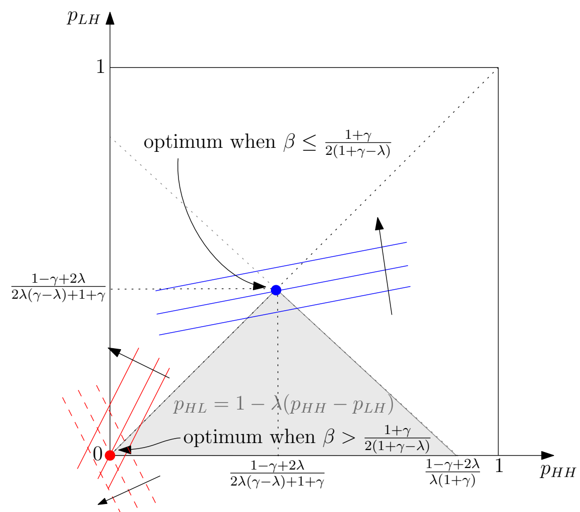


Figure 4: The optimal values of (p_{HH}, p_{LH}) when $\lambda \le (1+\gamma)/2$ given $p_{HL} = 1-\lambda(p_{HH}-p_{LH})$. *Note*: In the figure, $(1-\gamma+2\lambda) > 0$ and $\gamma > 1$ (so that $\frac{1-\gamma+2\lambda}{\lambda(1+\gamma)} < 1$) are assumed.

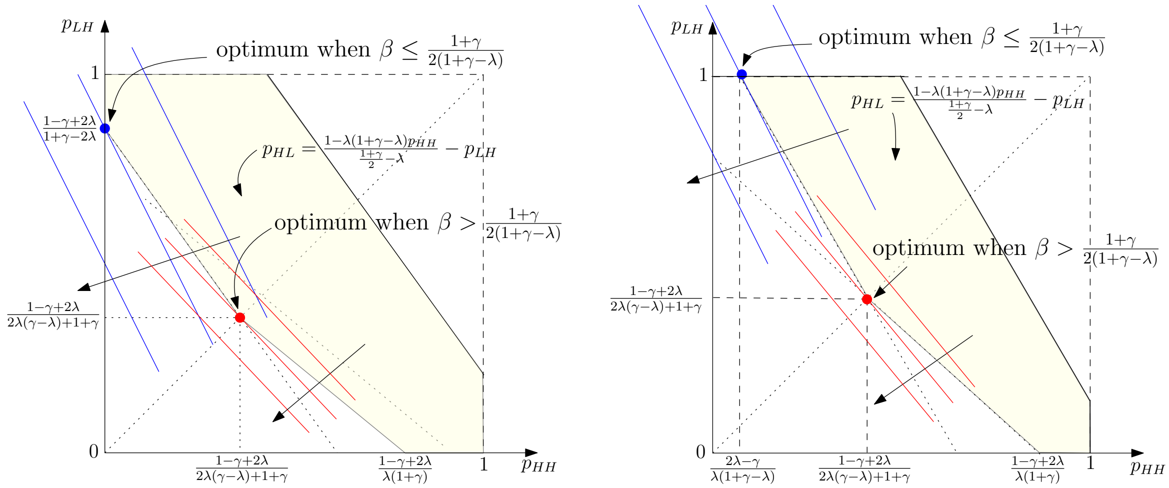
where the first inequality follows from $p_{HL} \ge 0$, the second inequality follows from $p_{HL} \le 1$, and the third inequality is (IC_{1H}). Recall from Remark 3 that both coefficients on p_{HH} and p_{LH} in the objective function $W_\beta^{(3)}$ are strictly negative. Thus decreasing both p_{HH} and p_{LH} as much as possible is optimal.

(i) For $\beta \le \hat{\beta}$, it can easily be verified that the MRS of p_{LH} for p_{HH}, $\frac{\lambda\left(\frac{\beta(1+\gamma-\lambda)}{(1+\gamma)/2-\lambda}-1\right)}{2\beta-1}$, is greater than the absolute value of the coefficient on p_{HH} in the first (or second) inequality above, $\frac{2\lambda(1+\gamma-\lambda)}{1+\gamma-2\lambda}$. (Note that this is always greater than $\frac{\lambda(1+\gamma)}{(\lambda+1)(1+\gamma-2\lambda)}$ for $\lambda < \gamma$.) Thus if $\lambda > \gamma/2$, then the optimum occurs at $p_{LH} = 1$ and $p_{HH} = \frac{2\lambda-\gamma}{\lambda(1+\gamma-\lambda)} \in (0,1)$ in which case $p_{HL} = 1$; and if $\lambda \le \gamma/2$, then the optimum occurs at $p_{LH} = \max\{0, \frac{1-\gamma+2\lambda}{1+\gamma-2\lambda}\} \le 1$ and $p_{HH} = 0$ in which case $p_{HL} = \min\{1, \frac{1}{(1+\gamma)/2-\lambda}\} \le 1$.

(ii) For $\beta > \hat{\beta}$, it can also be verified that $\frac{\lambda(1+\gamma)}{(\lambda+1)(1+\gamma-2\lambda)} < \frac{\lambda\left(\frac{\beta(1+\gamma-\lambda)}{(1+\gamma)/2-\lambda}-1\right)}{2\beta-1} < \frac{2\lambda(1+\gamma-\lambda)}{1+\gamma-2\lambda}$. Thus the optimum occurs at $p_{HH} = p_{LH} = \max\{0, \frac{1-\gamma+2\lambda}{2\lambda(\gamma-\lambda)+1+\gamma}\} < 1$.

Figure 5 illustrates the optimal values of (p_{HH}, p_{LH}) conditional on $p_{HL} = \frac{1-\lambda(1+\gamma-\lambda)p_{HH}}{(1+\gamma)/2-\lambda} - p_{LH}$.

We now compare the values of the objective function $W_\beta^{(k)}$, $k = 1, 2, 3$, in Remark 3 by plugging in the optimal values of p_{HH} and p_{LH} for each given p_{HL}, and determine the optimal



(a) $\lambda \leq \frac{\gamma}{2}$ (b) $\lambda > \frac{\gamma}{2}$

Figure 5: The optimal values of (p_{HH}, p_{LH}) when $\lambda < (1+\gamma)/2$ given $p_{HL} = \frac{1-\lambda(1+\gamma-\lambda)p_{HH}}{(1+\gamma)/2-\lambda} - p_{LH}$. *Note*: In the figure, $(1-\gamma+2\lambda) > 0$ and $\gamma > 1$ (so that $\frac{1-\gamma+2\lambda}{\lambda(1+\gamma)} < 1$) are assumed.

solution $(p_t)_{t\in T}$ that solves **(RP)**. We then characterize the optimal solution $(p_t, B_t)_{t\in T}$ in the complete (unrelaxed) mediation programme for case **(A)**; in doing so, we check that the solution of **(RP)** satisfies the omitted or relaxed constraints and show the slacks of non-binding constraints.

When $\beta > \hat{\beta}$: If $1-\gamma+2\lambda > 0$, then $W_\beta^{(1)} = W_\beta^{(2)} = \beta(\lambda+1)$, and $W_\beta^{(3)} = \beta(\lambda+1) + [\beta\lambda - (2\beta-1)(1+\gamma)/2](\lambda+1)p_{HH} < \beta(\lambda+1)$. If $1-\gamma+2\lambda \leq 0$, then the only possible value of p_{HL} equals $\frac{1-\lambda(1+\gamma-\lambda)p_{HH}}{(1+\gamma)/2-\lambda} - p_{LH}$ in which case $p_{HH} = p_{LH} = 0$ and $p_{LL} = 1$, and so $W_\beta^{(3)} = \beta\frac{1+\gamma}{1+\gamma-2\lambda}$ is the maximal objective function. Thus the solution in the relaxed programme **(RP)** consists of the following:

(a) For $\lambda \leq (\gamma-1)/2$ (given $\gamma > 1$): $p_{LL} = 1$, $p_{HL} = \frac{1}{(1+\gamma)/2-\lambda} \in (0,1]$, $p_{LH} = p_{HH} = 0$. Plugging in the binding constraints, we have:

$$(IC_{1L}) \quad B_{LL} = \frac{1}{(1+\gamma)/2-\lambda}(B_{HL} + (1+\gamma)/2) = \frac{1+\gamma}{1+\gamma-2\lambda} > 0;$$

$$(IC_{2L}) \quad \frac{\lambda}{(1+\gamma)/2-\lambda}(1-B_{HL}) + (1-B_{LL}) = 0;$$

$$(PC_{2H}) \quad 0 = 0,$$

where $\frac{1}{(1+\gamma)/2-\lambda}B_{HL} = 0$ by the binding (PC_{1H}). The information rents are $du_{1H} = \frac{1}{(1+\gamma)/2-\lambda}B_{HL} = 0$ by the binding (PC_{1H}); $du_{1L} = B_{LL} = \frac{1+\gamma}{1+\gamma-2\lambda} > 0$ by the binding (IC_{1L}) and (PC_{1H}); $du_{2H} = 0$ by the binding (PC_{2H}); and $du_{2L} = \frac{\lambda}{(1+\gamma)/2-\lambda}(1-B_{HL}) + (1-B_{LL}) = 0$ by the binding (IC_{2L}). The probability constraints are trivially satisfied for the unrelaxed programme. We verify the slack of the omitted constraint (IC_{2H}) is $\frac{\lambda}{(1+\gamma)/2-\lambda} + 1 > 0$. We verify Lemma 3 that (PC_{1L}) is satisfied with slack, $B_{LL} > 0$; and (PC_{2L}) is satisfied with equality, $\frac{\lambda}{(1+\gamma)/2-\lambda}(1 - B_{HL}) + (1 - B_{LL}) = 0$ (by the binding (IC_{2L})). In the solution of the unrelaxed programme, the binding constraints are (IC_{1L}), (IC_{2L}), (PC_{1H}), (PC_{2H}), and (PC_{2L}).

(b) For $(\gamma - 1)/2 < \lambda$: $p_{LL} = p_{HL} = 1$ and $p_{HH} = p_{LH} = 0$. Plugging in the binding (IC_{1L}) and (IC_{2L}) constraints, $B_{LL} = B_{HL} + (1+\gamma)/2$ and $\lambda(1 - B_{HL}) + 1 - B_{LL} = 0$, we obtain $B_{HL} = 1 - \frac{1+\gamma}{2(\lambda+1)} \in (0, 1)$ and $B_{LL} = 1 + \frac{\lambda(1+\gamma)}{2(\lambda+1)} > 0$, and so the information rents (or interim payoffs) are $du_{1H} = B_{HL}$, $du_{1L} = B_{LL}$, $du_{2H} = 0$, and $du_{2L} = \lambda(1 - B_{HL}) + (1 - B_{LL}) = 0$. The probability constraints are trivially satisfied for the unrelaxed programme. We verify that the slack of the omitted constraint (IC_{2H}) is $\lambda+1 > 0$. We verify Lemma 3 that (PC_{1L}) is satisfied with slack, $B_{LL} > 0$; and (PC_{2L}) is satisfied with equality, $\lambda(1 - B_{HL}) + (1 - B_{LL}) = 0$ (by the binding (IC_{2L})). In the solution of the unrelaxed programme, the binding constraints are (IC_{1H}), (IC_{1L}), (IC_{2L}), (PC_{2H}), and (PC_{2L}).

When $\beta \le \hat{\beta}$: For $\lambda \le \gamma/2$, $W_\beta^{(1)} = W_\beta^{(3)} = \beta(\lambda + 1) + [\beta\lambda - (2\beta - 1)(1 + \gamma)/2]\frac{1-\gamma+2\lambda}{1+\gamma-2\lambda} > W_\beta^{(2)} = \beta(\lambda + 1) + [\beta\lambda - (2\beta - 1)(1 + \gamma)/2]\frac{(\lambda+1)(1-\gamma+2\lambda)}{2\lambda(\gamma-\lambda)+1+\gamma}$, where the strict inequality follows from $2\lambda(\gamma - \lambda) + 1 + \gamma > (\lambda + 1)(1 + \gamma - 2\lambda)$ which trivially holds assuming $1 - \gamma + 2\lambda > 0$. Note that if $1 - \gamma + 2\lambda \le 0$, then the only possible value of p_{HL} equals $\frac{1-\lambda(1+\gamma-\lambda)p_{HH}}{(1+\gamma)/2-\lambda} - p_{LH}$ in which case $p_{HH} = p_{LH} = 0$ and $p_{LL} = 1$, and so $W_\beta^{(3)} = \beta\frac{1+\gamma}{1+\gamma-2\lambda}$ is the maximal objective function. For $\lambda > \gamma/2$, $W_\beta^{(1)} = W_\beta^{(3)} = \beta(\lambda + 1) + [\beta\lambda - (2\beta - 1)(1 + \gamma)/2]\frac{\lambda+1}{1+\gamma-\lambda} > W_\beta^{(2)} = \beta(\lambda + 1) + [\beta\lambda - (2\beta - 1)(1 + \gamma)/2]\frac{(\lambda+1)(1-\gamma+2\lambda)}{2\lambda(\gamma-\lambda)+1+\gamma}$, where the strict inequality follows from $2\lambda(\gamma - \lambda) + 1 + \gamma > (1 + \gamma - \lambda)(1 - \gamma + 2\lambda)$ which holds for $\lambda < \gamma$. Thus the solution in the relaxed programme **(RP)** consists of the following:

(a) For $\lambda \leq (\gamma-1)/2$ (given $\gamma > 1$): $p_{LL} = 1$, $p_{HL} = \frac{1}{(1+\gamma)/2-\lambda} \in (0,1]$, $p_{LH} = p_{HH} = 0$. The characterization is exactly the same as in the case of (1)(a) above.

(b) For $(\gamma-1)/2 < \lambda \leq \gamma/2$: $p_{HL} = 1$, $p_{LH} = \frac{1-\gamma+2\lambda}{1+\gamma-2\lambda} \in (0,1)$, and $p_{HH} = 0$. Plugging in the binding constraints, we have:

$$
\begin{aligned}
(IC_{1L}) \quad & \frac{\lambda(1-\gamma+2\lambda)}{1+\gamma-2\lambda}B_{LH} + B_{LL} = B_{HL} + (1+\gamma)/2 = (1+\gamma)/2 > 0; \\
(IC_{2L}) \quad & \lambda(1-B_{HL}) + (1-B_{LL}) = \left(\frac{1-\gamma+2\lambda}{1+\gamma-2\lambda}\right)(1+\gamma)/2 > 0; \\
(PC_{2H}) \quad & \frac{1-\gamma+2\lambda}{1+\gamma-2\lambda}(1-B_{LH}) = 0,
\end{aligned}
$$

where $B_{HL} = 0$ by the binding (PC_{1H}). The information rents are $du_{1H} = B_{HL} = 0$ by the binding (PC_{1H}); $du_{1L} = \frac{\lambda(1-\gamma+2\lambda)}{1+\gamma-2\lambda}B_{LH} + B_{LL} = (1+\gamma)/2 > 0$ by the binding (IC_{1L}) and (PC_{1H}); $du_{2H} = \frac{1-\gamma+2\lambda}{1+\gamma-2\lambda}(1-B_{LH}) = 0$ by the binding (PC_{2H}); and $du_{2L} = \lambda(1-B_{HL}) + (1-B_{LL}) = \frac{(1+\gamma)(1-\gamma+2\lambda)}{2(1+\gamma-2\lambda)} > 0$ by the binding (IC_{2L}). That is, $B_{HL} = 0$, $B_{LH} = 1$, $B_{LL} = \frac{((1+\gamma)/2)^2-\lambda(\lambda+1)}{(1+\gamma)/2-\lambda}$. The probability constraints are trivially satisfied for the unrelaxed programme. We verify the slack of the omitted constraint (IC_{2H}) is $\lambda+1 > \frac{1-\gamma+2\lambda}{1+\gamma-2\lambda}$ because $\lambda+1 > 1$, and $\frac{1-\gamma+2\lambda}{1+\gamma-2\lambda} \leq 1$ for $\lambda \leq \gamma/2$. We verify Lemma 3 that the slacks of (PC_{1L}) and (PC_{2L}) are, respectively, $\frac{\lambda(1-\gamma+2\lambda)}{1+\gamma-2\lambda}B_{LH} + B_{LL} > 0$ and $\lambda(1-B_{HL}) + (1-B_{LL}) > 0$. In the solution of the unrelaxed programme, the binding constraints are (IC_{1L}), (IC_{2L}), (PC_{1H}), and (PC_{2H}).

(c) For $\lambda > \gamma/2$: $p_{LL} = p_{HL} = p_{LH} = 1$ and $p_{HH} = \frac{2\lambda-\gamma}{\lambda(1+\gamma-\lambda)} \in (0,1)$. Plugging in the binding constraints, we have:

$$
\begin{aligned}
(IC_{1L}) \quad & \lambda B_{LH} + B_{LL} = \frac{2\lambda-\gamma}{1+\gamma-\lambda}(B_{HH} + (1+\gamma)/2) + (B_{HL} + (1+\gamma)/2); \\
(IC_{2L}) \quad & \lambda(1-B_{HL}) + (1-B_{LL}) = \left(\frac{\lambda+1}{1+\gamma-\lambda}\right)(1+\gamma)/2 > 0; \\
(PC_{2H}) \quad & \frac{2\lambda-\gamma}{1+\gamma-\lambda}(1-B_{HH}) + (1-B_{LH}) = 0.
\end{aligned}
$$

The information rents are $du_{1H} = \frac{2\lambda-\gamma}{1+\gamma-\lambda}B_{HH} + B_{HL} = 0$ by the binding (PC_{1H});

$du_{1L} = \lambda B_{LH} + B_{LL} = \frac{(\lambda+1)(1+\gamma)}{2(1+\gamma-\lambda)} > 0$ by the binding (IC_{1L}) and (PC_{1H}); $du_{2H} = \frac{2\lambda-\gamma}{1+\gamma-\lambda}(1 - B_{HH}) + (1 - B_{LH}) = 0$ by the binding (PC_{2H}); and $du_{2L} = \lambda(1 - B_{HL}) + (1 - B_{LL}) = \frac{(\lambda+1)(1+\gamma)}{2(1+\gamma-\lambda)} > 0$ by the binding (IC_{2L}). The probability constraints are trivially satisfied for the unrelaxed programme. We verify that the slack of the omitted constraint (IC_{2H}) is $\lambda + 1 > (\lambda + 1)/(1 + \gamma - \lambda)$. We verify Lemma 3 that the slacks of (PC_{1L}) and (PC_{2L}) are, respectively, $\lambda B_{LH} + B_{LL} = \frac{(\lambda+1)(1+\gamma)}{2(1+\gamma-\lambda)} > 0$, and $\lambda(1 - B_{HL}) + (1 - B_{LL}) = \frac{(\lambda+1)(1+\gamma)}{2(1+\gamma-\lambda)} > 0$. In the solution of the unrelaxed programme, the binding constraints are (IC_{1L}), (IC_{2L}), (PC_{1H}), and (PC_{2H}).

Case **(B)** $(1 + \gamma)/2 < \lambda < \gamma$: The maximal possible value of p_{HL} in the relaxed problem **(RP1)** is determined by $p_{HL} = \min\{1, 1 - \lambda(p_{HH} - p_{LH})\}$ under the condition that $p_{HL} \geq \frac{1-\lambda(1+\gamma-\lambda)p_{HH}}{(1+\gamma)/2-\lambda} - p_{LH}$.

(1) If $p_{HL} = 1$, then the feasible set for (p_{HH}, p_{LH}) is defined by the following inequalities from (IC_{1H}) and (PC_{1H}):

$$\begin{aligned} p_{LH} &\geq p_{HH}, \\ p_{LH} &\geq -\frac{2\lambda(1+\gamma-\lambda)}{1+\gamma-2\lambda}p_{HH} + \frac{1-\gamma+2\lambda}{1+\gamma-2\lambda}, \end{aligned}$$

where $-\frac{2\lambda(1+\gamma-\lambda)}{1+\gamma-2\lambda} > 1$ and $\frac{1-\gamma+2\lambda}{1+\gamma-2\lambda} < 0$ for $\lambda \in ((1 + \gamma)/2, \gamma)$.

(2) If $p_{HL} = 1 - \lambda(p_{HH} - p_{LH})$, then the (IC_{1H}) constraint binds and the feasible set for (p_{HH}, p_{LH}) is characterized by:

$$\begin{aligned} p_{LH} &\leq p_{HH}, \\ p_{LH}(\lambda + 1) &\geq -\frac{\lambda(1+\gamma)}{1+\gamma-2\lambda}p_{HH} + \frac{1-\gamma+2\lambda}{1+\gamma-2\lambda}, \end{aligned}$$

where the first inequality follows from $p_{HL} \leq 1$ and the binding (IC_{1L}), and the second inequality is (PC_{1H}). Note that $-\frac{2\lambda(1+\gamma-\lambda)}{1+\gamma-2\lambda} > -\frac{\lambda(1+\gamma)}{(\lambda+1)(1+\gamma-2\lambda)} > 1$ for for $\lambda \in ((1 + \gamma)/2, \gamma)$.

The feasible set for (p_{HH}, p_{LH}) for each possible p_{HL} is illustrated in Figure 6.

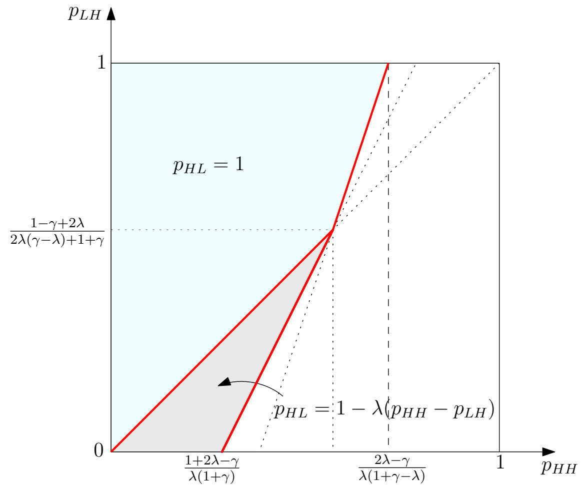
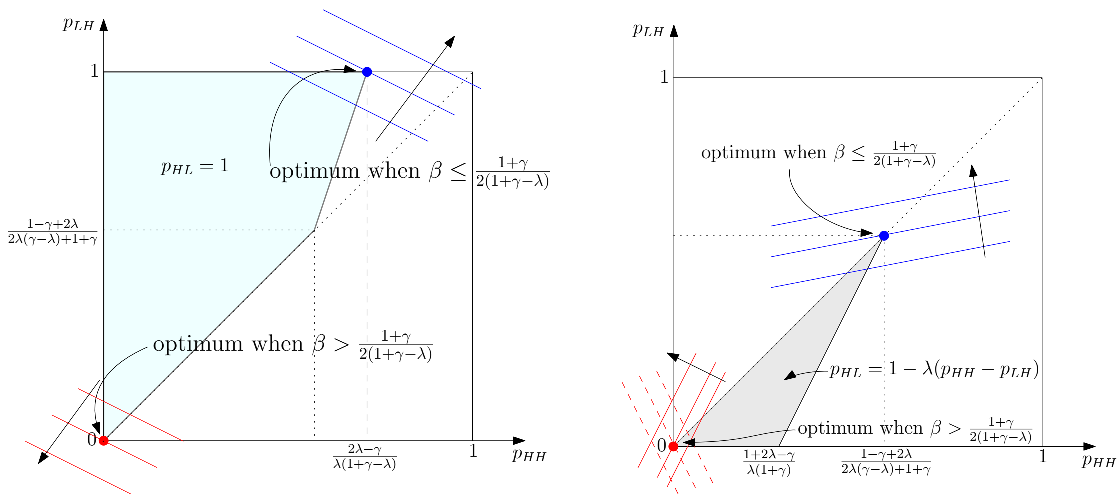


Figure 6: Feasible set for (p_{HH}, p_{LH}) when $\lambda > (1+\gamma)/2$

For each given p_{HL}, the values of p_{HH} and p_{LH} that maximize W_β in **(RP)** can be found in the same way as in Case **(A)** (1) and (2), which is illustrated in Figure 7; so we do not repeat the mathematical proof.



(a) Given $p_{HL} = 1$ (b) Given $p_{HL} = 1 - \lambda(p_{HH} - p_{LH})$

Figure 7: The optimal values of (p_{HH}, p_{LH}) when $\lambda > (1+\gamma)/2$

By comparing the values of the objective function $W_\beta^{(k)}$, $k = 1, 2$, given the optimal values of p_{HH} and p_{LH} for each given p_{HL}, we can determine the optimal solution $(p_t)_{t \in T}$ that solves **(RP)**. Thus the optimal mediation programme solution $(p_t, B_t)_{t \in T}$ for case **(B)** is characterized as follows: When $\beta > \hat{\beta}$: $W_\beta^{(1)} = W_\beta^{(2)} = \beta(\lambda + 1)$. Thus $p_{HH} = p_{LH} = 0$

and $p_{LL} = p_{HL} = 1$ in the relaxed programme **(RP)**. This exactly corresponds to the optimal solution for case **(A)** when $\beta > \hat{\beta}$, which solves the unrelaxed programme. When $\beta \leq \hat{\beta}$: $W_\beta^{(1)} > W_\beta^{(2)}$. Thus the solution in the relaxed programme **(RP)** consists of $p_{LL} = p_{HL} = p_{LH} = 1$ and $p_{HH} = \frac{2\lambda-\gamma}{\lambda(1+\gamma-\lambda)} \in (0, 1)$. This coincides with the optimal solution for case **(A)** when $\beta \leq \hat{\beta}$ for $\lambda > \gamma/2$, which solves the unrelaxed programme.

This completes the proof of Proposition 1.

Appendix D Proofs of Propositions 2 and 3

Proof of Proposition 2. We prove for the case of $\beta \leq 1/2$; the proof for the case of $\beta \geq 1/2$ is exactly analogous by interchanging the roles of players 1 and 2. Given a β-mediator with $\beta \leq 1/2$, suppose that all types of all players accept, i.e., $v_i(t_i) = 1$, for all i and t_i.

We first consider a deviation by player 1 at the selection stage. If player 1 (unexpectedly) deviates by rejecting the given β-mediator, then player 1 does not change her own belief on player 2's type, summarized by q; but player 2 updates his belief about player 1's type being H to q_1^n, which is arbitrary. (Note that there is no consistency requirement for the off the equilibrium path rejection beliefs.) After rejection by player 1 alone, the default game is played with the (revised) beliefs $\{q_1 = q_1^n, q_2 = q\}$, the peace-maximizing equilibrium of which can be found using Proposition 4. Let $du_{it_i}^d$ denote the deviation rent-payoff of player i of type t_i in the peace-maximizing equilibrium of the default game given the updated beliefs $q_1 = q_1^n$ and $q_2 = q$. The L-type of both players will choose peace, regardless of their off equilibrium beliefs. Player 1 of type H, given her belief $q < (1-q)\gamma$ (or $\lambda < \gamma$), will optimally choose war in the default game, thus receiving $du_{1H}^d = 0$. Because player 1 of type H chooses war with probability one, announcing war with probability one is also the best response for player 2 of type H regardless of his belief q_1^n (see Lemma 2). Therefore, the payoff for player 1 of type L is $q(1-p)\theta + (1-q)(1/2)$, thus $du_{1L}^d = 1/2$.

(i) Suppose that the given β-mediator (unanimously accepted) is moderately biased against player 1, i.e., $\beta \in \left[1 - \frac{1+\gamma}{2(1+\gamma-\lambda)}, 1/2\right)$. Because the equilibrium (mediation) rent is $du_{1H} = 0 = du_{1H}^d$, H-type player 1 does not gain by deviating (rejecting). For $\lambda > \gamma/2$, the equilibrium rent for player 1 of type L is $du_{1L} = \frac{(1+\gamma)(\lambda+1)}{2(1+\gamma-\lambda)}$, which is strictly greater than $du_{1L}^d = 1/2$; so deviation is never profitable. For $(\gamma-1)/2 < \lambda \leq \gamma/2$, $du_{1L} = \frac{1-\gamma+2\lambda}{1+\gamma-2\lambda}(1+\gamma)/2$, in which case we have $du_{1L} \geq du_{1L}^d = 1/2$ if and only if $\lambda \geq \frac{\gamma(1+\gamma)}{2(2+\gamma)}$; so deviation is profitable when $\lambda < \frac{\gamma(1+\gamma)}{2(2+\gamma)}$. For $\lambda \leq (\gamma-1)/2$, the equilibrium rent for player 1 of type L is $du_{1L} = 0 < 1/2 = du_{1L}^d$; so deviation is always profitable. Note that $(\gamma-1)/2 < \frac{\gamma(1+\gamma)}{2(2+\gamma)} < \gamma/2$. Thus player 1 of type L has an incentive to deviate from unanimous acceptance of β-mediator who is moderately biased against player 1 if $\lambda < \frac{\gamma(1+\gamma)}{2(2+\gamma)}$. Otherwise ($\lambda \geq \frac{\gamma(1+\gamma)}{2(2+\gamma)}$), unanimous acceptance of β-mediator for any given $\beta \in \left[1-\hat{\beta}, 1/2\right)$ constitutes an equilibrium.

(ii) Suppose that the given β-mediator (unanimously accepted) is unbiased, i.e., $\beta = 1/2$. Again, H-type player 1 has no incentive to deviate because the equilibrium mediation rent is $du_{1H} = 0 = du_{1H}^d$. For $\lambda > \gamma/2$, $du_{1L} = \frac{(1+\gamma)(\lambda+1)}{2(1+\gamma-\lambda)} > 1/2 = du_{1L}^d$; for $\lambda \leq \gamma/2$, $du_{1L} = \frac{1+\gamma}{2(1+\gamma-2\lambda)} > 1/2 = du_{1L}^d$. Thus an L-type player 1 also has no incentive to deviate.

(iii) Suppose that the given β-mediator (unanimously accepted) is strongly biased against player 1, i.e., $\beta < 1 - \frac{1+\gamma}{2(1+\gamma-\lambda)}$. The equilibrium mediation rents are $du_{1H} = du_{1L} = 0$. Thus an L-type payer 1 would always deviate.

We now consider a deviation by player 2 at the selection stage. Player 1 updates her belief about player 2's type being H to q_2^n, which is arbitrary; but player 2 does not update. The default game is played with the (revised) beliefs $\{q_1 = q, q_2 = q_2^n$, in the same way as before. Thus the deviation rent-payoffs are $du_{2H}^d = 0$ and $du_{2L}^d = 1/2$. We can easily check that the equilibrium mediation payoffs for player 2 are always (weakly) greater than the deviation payoffs. That is, the favored player never has an incentive to deviate.

Thus unanimous acceptance of the unbiased mediator ($\beta = 1/2$) always constitutes an equilibrium (for $\lambda < \gamma$); unanimous acceptance of strongly biased mediators can never be sustained in equilibrium; and unanimous acceptance of moderately biased mediators can be sustained in equilibrium iff $\lambda \geq \frac{\gamma(1+\gamma)}{2(2+\gamma)}$. □

Proof of Proposition 3. Given the prior beliefs $q_1 = q_2 = q$, the ex ante probability of peace in the β-mediator's optimal mechanism is computed as $q^2 p_{HH} + q(1-q)p_{HL} + (1-q)qp_{LH} + (1-q)^2 p_{LL}$. From Lemma 1 and Proposition 1, we can compute the ex ante chance of peace for any β-mediator with $\beta \in [1-\hat{\beta}, \hat{\beta}]$ to be $q(1-q)\frac{2}{1+\gamma-2\lambda} + (1-q)^2$ when $\lambda \leq \gamma/2$ and $q^2\frac{2\lambda-\gamma}{\lambda(1+\gamma-\lambda)} + 2q(1-q) + (1-q)^2$ when $\lambda > \gamma/2$. (Note that the optimal mechanisms of unbiased mediator and moderately biased mediators are identical when $\lambda > \gamma/2$. When $\lambda \leq \gamma/2$, the optimal recommendation probabilities of unbiased mediator and moderately biased mediators differ, but at the optimum, the ex ante probabilities of recommending peaceful splits to asymmetric dyads are equivalent to $q(1-q)\frac{2}{1+\gamma-2\lambda}$.) For any β-mediator with $\beta > \hat{\beta}$ or $\beta < 1-\hat{\beta}$, the ex ante chance of peace is $q(1-q)\frac{2}{1+\gamma-2\lambda} + (1-q)^2$ when $\lambda \leq (\gamma-1)/2$ given $\gamma > 1$, and is $q(1-q) + (1-q)^2$ when $\lambda > (\gamma-1)/2$. We can easily verify that $q(1-q) + (1-q)^2 < q(1-q)\frac{2}{1+\gamma-2\lambda} + (1-q)^2$ when $(\gamma-1)/2 < \lambda \leq \gamma/2$, and $q(1-q)+(1-q)^2 < q^2\frac{2\lambda-\gamma}{\lambda(1+\gamma-\lambda)} + 2q(1-q) + (1-q)^2$ when $\lambda > \gamma/2$. Hence, if $\lambda \leq (\gamma-1)/2$ and $\gamma > 1$, then the optimal mechanism of any strongly biased mediator achieves the same ex ante chance of peace as the unbiased (or any moderately biased) mediator; otherwise (if $\lambda > (\gamma-1)/2$), it achieves a strictly lower ex ante chance of peace. □

References

Ambrus, Attila, Eduardo M. Azevedo, Yuichiro Kamada and Yuki Takagi. 2013. "Legislative Committees as Information Intermediaries: A Unified Theory of Committee Selection and Amendment Rules." *Journal of Economics Behavior & Organization* 94:103–115.

Beber, Bernd. 2012. "International Mediation, Selection Effects, and the Question of Bias." *Conflict Management and Peace Science* 29(4):397–424.

Bercovitch, Jacob and Scott Sigmund Gartner. 2008. *International Conflict Mediation: New Approaches and Findings.* Routledge.

Bester, Helmut and Karl Wärneryd. 2006. "Conflict and the Social Contract." *Scandinavian Journal of Economics* 108(2):231–49.

Celik, Gorkem and Michael Peters. 2011. "Equilibrium Rejection of a Mechanism." *Games and Economic Behavior* 73:375–387.

Cramton, Peter C. and Thomas R. Palfrey. 1995. "Ratifiable Mechanisms: Learning from Disagreement." *Games and Economic Behavior* 10:255–283.

Fey, Mark and Kristopher W. Ramsay. 2009. "Mechanism Design Goes to War: Peaceful Outcomes with Interdependent and Correlated Types." *Review of Economic Design* 13(3):233–250.

Fey, Mark and Kristopher W. Ramsay. 2010. "When Is Shuttle Diplomacy Worth the Commute? Information Sharing through Mediation." *World Politics* 62(04):529–560.

Fisher, Ronald J. 1995. Pacific, Impartial Third-Party Intervention in International Conflict: A Review and Analysis. In *Beyond Confrontation:Learning Conflict Resolution in the Post-Cold War Era*, ed. John A Vasquez, James Turner Johnson, Sanford Jaffe and Linda Stamato. Ann Arbor: University of Michigan Press pp. 39–59.

Frazier, Derrick V. and William J Dixon. 2006. "Third-Party Intermediaries and Negotiated Settlements, 1946–2000." *International Interactions* 32(4):385–408.

Hörner, Johannes, Massimo Morelli and Francesco Squintani. 2015. "Mediation and Peace." *The Review of Economic Studies* 82(4):1483–1501.

Izumi, Carol. 2010. "Implicit Bias and the Illusion of Mediator Neutrality." *Washington University Journal of Law & Policy* 34.

Kleiboer, Marieke. 1996. "Understanding Success and Failure of International Mediation." *Journal of Conflict Resolution* 40(2):360–389.

Kydd, Andrew. 2003. "Which Side Are You On? Bias, Credibility, and Mediation." *American Journal of Political Science* 47(4):597–611.

Rauchhaus, Robert W. 2006. "Asymmetric Information, Mediation, and Conflict Management." *World Politics* 58(2):207–241.

Salamanca, Andrés. 2024. "Biased Mediators in Conflict Resolution." *American Law and Economics Review* Forthcoming.

Savun, Burcu. 2008. "Information, Bias, and Mediation Success." *International Studies Quarterly* 52(1):25–47.

Smith, William P. 1985. "Effectiveness of the Biased Mediator." *Negotiation Journal* 1(4):363–372.

Svensson, Isak. 2007. "Bargaining, Bias and Peace Brokers: How Rebels Commit to Peace." *Journal of Peace Research* 44(1):177–194.

Svensson, Isak. 2013. "Research on Bias in Mediation: Policy Implications." *Penn State Journal of Law & International Affairs* 2(1):17–26.

Tan, Guofu and Okan Yilankaya. 2007. "Ratifiability of Efficient Collusive Mechanisms in Second-Price Auctions with Participation Costs." *Games and Economic Behavior* 59:383–396.

Young, Oran R. 1967. *The Intermediaries: Third Parties in International Crises*. Princeton: Princeton University Press.