Toward an Understanding of Optimal Mediation Choice[*]

Jin Yeub Kim[†]   Wooyoung Lim[‡]

May 8, 2025

Abstract

Mediation is a strategic tool in mitigating conflict in bargaining with incomplete information. We experimentally investigate the informed principal problem of mediator selection. The theory of neutral optimum predicts that the principal's optimal inscrutable choice is not the one that maximizes the ex-ante probability of peace in our environment due to the conflicting interests of principal types. We find that subjects do not choose the neutral mediator more often than the peace-maximizing one. Different types of principal subjects, while acknowledging the need for inscrutable mediator selection, do not agree on their intertype compromise outlined by the theory of neutral optimum. This observed behavior aligns with predictions from projection bias, which may pose a significant behavioral obstacle to achieving intertype compromise under the neutral optimum.

Keywords: Informed Principal Problems, Mechanism Selection, Mediation, Inscrutability, Neutral Optimum, Laboratory Experiments

JEL classification: C72, C91, D82

[*] We received helpful feedback from Ala Avoyan, Andreas Blume, Vincent Crawford, John Duffy, Alex Frankel, John List, Mike McBride, Roger Myerson, Takeshi Nishimura, Daniela Puzzello, Andrés Salamanca, Lars Stole, and seminar participants at Yonsei Micro Reading Group seminar, UCI Experimental Group seminar, UChicago Experimental Group seminar, the 2024 Asia-Pacific ESA meeting in Singapore, the 2024 Conference on Mechanism and Institution Design (CMID) in Budapest, and the 2024 SUSTech-SZU Micro Workshop in Shenzhen. We thank Ho Fung Leung, Zhongying Hu, and Zhong Zhang for their research assistance. This study is supported by a grant from the Research Grants Council of Hong Kong (Grant No. GRF16504523).

[†] Associate Professor, School of Economics, Yonsei University, 50 Yonsei-ro Seodaemun-gu, Seoul 03722, Republic of Korea, E-mail: jinyeub@yonsei.ac.kr

[‡] Associate Professor, Department of Economics, The Hong Kong University of Science and Technology, Clearwater Bay, Kowloon, Hong Kong, E-mail: wooyoung@ust.hk

1 Introduction

Mediation is one of the most commonly used methods of third-party intervention for dispute resolution in many kinds of conflicts, ranging from labor-management and legal disputes to international conflicts.[1] When the source of these conflicts is asymmetric information between disputants, mediation can be used to reduce information asymmetries and thus to minimize the risk of conflict. Given the importance of mediation, a large academic literature examines the effectiveness of mediation and the impact of various features of mediation on conflict resolution (see, e.g., Fey and Ramsay, 2009, 2010; Goltsman et al., 2009; Salamanca, 2024; Wilkenfeld et al., 2003, among many others).

Most theoretical works on mediation applying the mechanism design approach take mediators as exogenous and assume that the mediator's objective is the minimization of the ex ante probability of conflict (see, e.g., Bester and Wärneryd, 2006; Hörner, Morelli and Squintani, 2015). Such an assumption might be natural given that disputants seek the assistance of a mediator precisely as a means for reducing potential conflicts. In practice, however, disputants often choose a mediator themselves to bring to the mediation table among many available mediators that may differ in their relative effectiveness in bringing peace.[2] The fundamental question is then what kinds of mediators should we expect to observe to be chosen more often by disputants?

The answer to this question is not obvious especially when the disputant with the authority to select a mediator has private information, whom we call the *informed principal*. On the one hand, the informed principal may want to choose the mediator that she most

[1] The use of alternative dispute resolution (ADR) techniques, such as arbitration and mediation, in civil trials have prominently increased since the implementation of the Civil Justice Reform Act of 1990 in the U.S. (Galanter, 2004). Stienstra (2011) reports that more than one-third of all federal trial courts authorise some forms of ADR, two thirds of which are mediation. For the types of legal cases where mediation can be used, see the Mediation Section on the American Bar Association's website, available at `https://www.americanbar.org/groups/public_education/resources/law_related_education_network/how_courts_work/mediation_whenuse` (accessed September 27, 2024). For incidences of mediation in international conflicts, see Bercovitch and Gartner (2009) and Frazier and Dixon (2006).

[2] Wilkenfeld et al. (2003) document using historical data and experimental approach that the effectiveness of mediation varies with mediator style in international crises. The American Arbitration Association offers mediation to parties in various disputes in industries and fields, and provides a list of mediators from which the parties can choose (see `https://www.aaamediation.org/find-a-mediator`, accessed September 27, 2024). Individual mediator profiles provide mediator details such as mediator experience, style, and process preferences, which would impact mediation success rates.

prefers. On the other hand, her mediator choice may reveal some information about herself that she wants to conceal. In this paper, we study how such a dilemma is resolved, that is, how the concern of mediator selection potentially leaking private information shapes the mediator choice.

The analysis of mechanism selection by an informed principal offers theoretical insights. In his seminal paper, Myerson (1983) develops a theory of inscrutable mechanism selection and proposes a variety of solution concepts that satisfy *inscrutability*: the informed principal should choose the mechanism that is a reasonable selection for all types of the principal, so that the selection itself conveys no information. If the principal's different types prefer different mediators, then for inscrutability she must select the one that will be perceived as a reasonable *intertype compromise* between the conflicting preferences of different possible types. One key solution that delimits this inscrutable intertype compromise is the concept of *neutral optimum*.[3]

We take the informed principal's mechanism selection game to the lab. To our knowledge, this is the first paper that experimentally analyzes the informed principal problem. To keep the analysis manageable and induce salient incentives, we use a simple conflict situation with a two-type information structure and two possible choices of mediators implemented by computer algorithms. The primary interest of our paper is in testing the concept of neutral optimum and the subjects' behavior of inscrutable selection in an experimental setting.

For our baseline environment, we adopt a simplified version of the standard conflict model studied in Hörner, Morelli and Squintani (2015). In the model and in the experiment, two players compete for a fixed amount of surplus. They can jointly opt for an agreement sharing the surplus equally, or either player can inflict disagreement which shrinks the surplus. Each player's type can be either high or low and is private information that determines the players' disagreement payoffs. The high type prefers agreement over disagreement only with the same-type opponent, whereas the low type always prefers agreement regardless of the opponent's type. In this setting, the players can use mediation, under which the players confidentially send messages about their types to a mediator with a commonly known algorithm; then the

[3] The neutral optimum is an axiomatically founded solution concept that cannot be blocked with any reasonable theory of blocking. See Myerson (1983) for the formal definition.

mediator chooses message-dependent prescriptions of agreement or disagreement.[4]

We consider situations in which an informed principal chooses a mediator among the two options of interim incentive efficient mediators, one of which we call the *peace-maximizing* mediator and the other the *neutral* mediator.[5] Pooling on either mediator can be supported as a sequential equilibrium of our mediator selection game (defined later), which can also be justified by the idea of inscrutable intertype compromise. The concept of neutral optimum (Myerson, 1983) predicts for our setting that pooling on the neutral mediator, which is preferred by the high type, is the most reasonable intertype compromise. On the other hand, the peace-maximizing mediator corresponds to the one that maximizes the ex ante chance of peace, which is preferred by the low type. It is the unique ex ante incentive efficient mediator in our setting, so the only optimal choice by an uninformed principal if she were to choose before learning her type.[6]

We conducted two experiments that were designed in a sequential manner. The first experiment aimed to collect observational data to test the theory of neutral optimum. After finding that the theory of neutral optimum did not work in the lab, the second experiment was designed to identify the primary source of the failure.

As the main treatments of our first experiment, we considered two versions of principal's mediator selection game depending on the information structure at the time of selection. In the *uninformed* mediator selection game, an uninformed principal chose a mediator and then each player learned their private type; in the *informed* mediator selection game, after the players were informed of their types, the informed principal made a choice. After this first stage of mediator selection in both versions, the other player (subordinate) was invited to report his belief about the principal's type based on the mediator choice. Finally, each player

[4] Hörner, Morelli and Squintani (2015) describe the mediator's recommendation (or prescription) of disagreement as a mediation (or arbitration) failure which leads to the players fighting a war. Casella, Friedman and Perez Archila (2024) describe it as the mediator's refusal to mediate (or walking out). From the perspective of the bargaining problem, the recommendation of disagreement should be taken as one possible recommendation of a negotiation outcome. In our setting, only one agreement recommendation (of an equal split) is possible, so the mediator with enforcement power and the one without coincide. The relevant discussions are given in more detail in Appendix B.1.

[5] The theory of efficient mechanisms (Holmström and Myerson, 1983) identifies either one as a plausible selection by an informed principal.

[6] The ex ante probability of peace is a natural measure of the ex ante efficiency for a large class of games, but ex ante efficiency is sensitive to utility normalizations (Ledyard and Palfrey, 1994).

reported its type to be used in playing the chosen mediator; at this stage, the subordinate may dispense with the chosen mediator instead of sending a report, inflicting disagreement.

We find that the peace-maximizing mediator is indeed chosen with a significantly higher proportion than the neutral mediator in the uninformed mediator selection game, in line with the theoretical prediction. However, in the informed mediator selection game, the neutral mediator is not chosen more often than the peace-maximizing mediator, contradicting the theory of neutral optimum. More importantly, we observe that the two principal types did not select the same mediator. This empirically-observed type-dependent choice may stem from subjects not understanding the benefit of inscrutable mediator choices. Alternatively, it could occur when both types recognize the benefit of inscrutable mediator choices so that they each internally execute the inscrutable intertype compromise, but struggle to reach an agreement on their intertype compromise.

To determine which scenario applies, we conducted the second experiment. In this experiment, the principals were first informed of their types. We then presented them with three options to elicit not only their choice of mediator given their informed type but also what would have been their choice if they were of the other type: (1) select the neutral mediator and choose the same one if they were of the other type (Neutral Compromising); (2) select the peace-maximizing mediator and choose the same one if they were of the other type (P-Max Compromising); and (3) select the preferred mediator among the two and choose the other one if they were of the other type (Uncompromising).[7] This design allows us to see what each informed principal believes the other type should do in making their inscrutable intertype compromise. Unlike in the strategy method (Selten, 1967), the informed principals made their choices after knowing their type, meaning that their chosen mediator in the unrealized-type scenario could not be implemented. As a result, each type of informed principal has in mind their interim utility maximization rather than ex ante utility maximization when navigating their intertype compromise.

Our experimental evidence indicates that *the theory of inscrutable mediator selection*

[7] We designed the experiment so that participants first engaged in twenty-four rounds of bargaining, eight each under each of the three mediator selection schemes, which are type-contingent rules for selecting a mediator (Parts 1-3). Following this, Part 4 began, where participants had the three options to indicate their informed type's choice of mediator as well as their choice if they were of the other type.

works but the concept of neutral optimum fails in the lab. We find that the subjects rarely chose the Uncompromising option and predominantly chose either one of the two Compromising options. This tendency suggests that the subjects understand the need for inscrutability and thus phrase their mediator choice in a way that is independent of their type, executing their inscrutable intertype compromise in some way. We also find that the subjects chose the P-Max Compromising option more often than the Neutral Compromising option, consistent with the mediator choice data in the first experiment. Our data further show that different types of principals chose different Compromising options, mostly seeking compromises that favored their own interests. These finding indicate that the subjects do not make compromises in the way that the concept of neutral optimum advocates.

The phenomenon in which each type favors a compromise that benefits its own type can be explained by *projection bias* (Loewenstein, O'Donoghue and Rabin, 2003). When high type principals consider a low type's maximization problem, projection bias leads them to mispredict the low type's interim utility, resulting in a conclusion that favors the high type. Conversely, when the realized type is low, the same bias leads the principal to draw an opposite conclusion. Our findings suggest that projection bias poses a significant behavioral obstacle to the intertype compromise, which Myerson's theory of inscrutable mechanism selection fails to address.

Furthermore, for different types to agree on the neutral optimum's intertype compromise, the subjects should recognize that when a principal's choice of peace-maximizing mediator revealed her type being low, a high type subordinate could easily exploit this information by rejecting the mediator. This would create strong pressure for the low type principals to resolve their compromise in favor of the high type. Importantly, the subjects should internally deliberate over these considerations *before* the selection process begins, and inscrutably choose the neutral mediator. If not, we might expect that subjects would learn over multiple rounds and incorporate these insights into their mediator choices; in particular, the low type principals might adjust their selection in subsequent rounds. Surprisingly, we found that high type subordinates did not predominantly decline the low type's preferred mediator when chosen. As a result, low type principals did not recognize that their preferred mediator choice could be detrimental, and thus felt no need to compromise by selecting the neutral mediator, which is the key behavioral explanation for the failure of the theory of neutral optimum in the lab.

Our paper contributes to the literature on informed principal problems. The problem of mechanism selection by an informed principal was pioneered by Myerson (1983), and developed by Maskin and Tirole (1990, 1992) taking a different approach from Myerson. A few authors study the problem in private value environments (Cella, 2008; Maskin and Tirole, 1990; Mylovanov and Tröger, 2012, 2014), and several others in common value environments (Balkenborg and Makris, 2015; Dosis, 2022; Koessler and Skreta, 2016, 2019; Myerson, 1983; Maskin and Tirole, 1992; Nishimura, 2022; Severinov, 2008; Skreta, 2011). We consider the common value case for our baseline model and employ the concept of neutral optimum to characterize a possible choice of mediator for subjects. Our contribution is an experimental test of Myerson's (1983) theory of inscrutable mechanism selection and neutral optimum.

This paper is also related to the few works on experimental tests of mechanism design and mediation. Blume, Lai and Lim (2023) experimentally compare mediated cheap-talk with direct cheap-talk communication. A more closely related experimental paper to ours is Casella, Friedman and Perez Archila (2024) who test the performance of the optimal mediation mechanism identified by Hörner, Morelli and Squintani (2015) over the optimal equilibrium of unmediated communication. Theory predicts that mediation can lead to a strictly higher frequency of peace than unmediated communication; contrary to the theory, Casella, Friedman and Perez Archila (2024) find that the frequency of peace is not higher. This finding is also confirmed by our experimental data as we have the subjects play unmediated communication as well. While the theoretical literature on the problem of informed principal's mechanism selection and on the effectiveness of mediation is quite large, the contributions are very few on the experimental side. To our knowledge, our paper is the first one to experimentally study the informed principal's mechanism selection in a mediation game.

The paper is organized as follows. In Section 2, we describe the baseline model, the mediator selection game and its theoretical predictions. In Section 3, we present our hypotheses and experimental parameterization. In Section 4, we describe the first experimental design and report our findings. In Section 5, we provide the second experimental design and its findings. In Section 6, we conclude and discuss possible extensions. Appendix A provides nonparametric test results. Additional discussions, theoretical characterizations, supplemental analyses for the experimental results are contained in the Online Appendix. The *Experimental Instructions* (url-linked) document, available on the authors' websites, provides the omitted descriptions of experimental procedures and the experimental instructions.

2 Theoretical Background

2.1 Conflict Environment

Our baseline model closely follows the model of conflict presented in Hörner, Morelli and Squintani (2015) (HMS) and used in Casella, Friedman and Perez Archila (2024) (CFP).

Two players (1 and 2) dispute a given surplus of size one. Each player can either choose to accept an equal split (agreement) or choose an outright war (disagreement). Each player can be of high (H) or low (L) type, privately and independently drawn from the same distribution with probability q and $1-q$ respectively. If the two players both choose to accept the equal split, then it is implemented and the players each receive half of the surplus regardless of their types. If either player chooses war, then war occurs, and the value of the surplus shrinks to $\theta < 1$ and is divided according to the two players' types: when the two players are of the same type, they have the same expected share of the remaining surplus, so each player's expected war payoff is $\theta/2$; when one player is H type and the other is L type, the H type player's expected share is $\delta > 1/2$ and the expected war payoff is $\delta\theta > 1/2$ while the L type player receives $(1-\delta)\theta$.[8] We restrict our attention to the set of parameters to exclude the possibility that both types prefer the equal split over war.[9]

In this setting, the players can communicate through a mediator who collects the players' private messages and makes recommendations of either an equal split or war. By the revelation principle, without loss of generality, the mediation game can be set up as a direct-revelation mechanism. We focus on mechanisms of the following form: After being informed of their own type, if both players agree to participate in the mediation, each player sends

[8] War can always be averted with the equal split $(1/2, 1/2)$ if $\delta\theta \leq 1/2$.

[9] When $q\theta/2 + (1-q)\delta\theta \leq 1/2$, the equal split $(1/2, 1/2)$ is preferable to war for both types, hence war can always be averted. As in HMS and CFP, we also assume $q\theta/2 + (1-q)\delta\theta > 1/2$, or $q < \frac{\delta\theta - 1/2}{\delta\theta - \theta/2}$. CFP use $\delta = 1$ as the experimental parameter, whereas we use $\delta = 0.8$ in our experiments.

a confidential message $m_i \in \{h, l\}$ to a mediator. Given reports $m = (m_1, m_2)$, the mediator recommends an equal split $(1/2, 1/2)$ with probability $p(m)$ or war with probability $1 - p(m)$. The war recommendation may represent the mediator's refusal to mediate, resulting in the players fighting a war; it is phrased in the lab as the mediator's "walking out" as used by CFP. We restrict attention to symmetric mechanisms, so we let $p_H \equiv p(h, h)$, $p_M \equiv p(h, l) = p(l, h)$, and $p_L \equiv p(l, l)$. A mediator commits to its mechanism (p_H, p_M, p_L); thus, we will use the words mediator and mechanism interchangeably throughout the paper.

Our baseline model and mediation protocol differ from those in HMS in that the only possible peaceful settlement is an equal split. Appendix B.1 provides the detailed descriptions and justifications for our choice of the baseline environment and the restriction.

2.2 Mediator Selection Game

We are interested in situations in which the players can choose their mediator among mediators that differ in mediation "styles" represented by recommendation probabilities. We call the player choosing a mediator the *principal* and the other player the *subordinate*. Our main objective is to experimentally investigate the theory of mechanism selection by an informed principal. To do so, we consider two cases regarding what information players possess at the time of selection: interim, when each player knows only her own type; and ex ante, before any player learns her type. We will refer to the former case as *informed mediator selection* and the latter as *uninformed mediator selection*, which serves as a benchmark.

For informed mediator selection, we modify Myerson's (1983, Sec. 5) mechanism selection game. The timing of the game is as follows.

1. Each player first learns her own type; then the informed principal selects and announces a mediator.
2. Each player confidentially reports its type to the principal's announced mediator, in which case the mediator's recommendation is implemented. But instead of sending a report, the subordinate can immediately choose to go to war.

The announcement of the principal's mediator choice may convey some information to the subordinate about the principal's type. In our setting, the principal's announced mediator cannot be implemented without the subordinate agreeing to participate. This is because war can be initiated unilaterally, and the subordinate can trigger war whenever it might be profitable given her information after the announcement.[10] We phrase this subordinate's action of going to war in the lab as *declining* the announced mediator. This option can be crucial for the subordinate because he may make some inferences about the principal's type based on the principal's announcement.

For uninformed mediator selection, we modify the above game so that in stage 1, an *uninformed* principal first selects and announces a mediator; then each player learns one's own type, after which the game proceeds in the same way.

2.3 Theoretical Predictions

We presuppose that the principal would choose among the *interim incentive efficient* (IIE) mechanisms (Holmström and Myerson, 1983).[11] A mechanism is feasible if it is incentive compatible (so that every player wishes to report her type truthfully) and individually rational (so that every player agrees to participate in the mediation), under the prior beliefs. A mechanism is IIE if it is feasible and it is not dominated by any other feasible mechanism.

There are infinitely many IIE mechanisms (or mediators) in our setting, and we provide theoretical predictions for the uninformed/informed principal's optimal mediator choice among those IIE mediators. The following two IIE mediators are of particular interest. One is associated with the highest ex ante probability of agreement (peace) among all IIE mediators, which we will label as "P-Max Mediator." The other one is the neutral optimum, a strong prediction made by Myerson (1983)'s theory of inscrutable mechanism selection, which we will label as "Neutral Mediator."[12]

[10] In stage 2, at the time the mediator is to be played, the subordinate decides concurrently whether to go to war or to participate in implementing the mediator, and in the latter case, whether to send a truthful report. The principal sends a report about her type to be used only in implementing the mediator. This setup is essentially due to the participation constraints that need to be satisfied in our environment. Appendix B.2 provides relevant details as well as justifications for our choice of the two-stage game rather than a three-stage game where the subordinate makes his participation and reporting decisions sequentially.

[11] In this paper, we do not discuss how we get incentive efficiency as a result of a mediator-selection process but rather take incentive efficiency as a requirement imposed on the set of available mediators. Maskin and Tirole (1992) characterize the set of equilibrium allocations of their mechanism selection game without requiring interim incentive efficiency. See Appendix B.4.

[12] Appendix B.3 formally characterizes and provides detailed descriptions of the set of IIE mediators and of the neutral optimum.

In uninformed mediator selection, principals choose a mediator before they know their type, so they would be concerned with their ex ante expected payoff in mediation. Therefore, it would be reasonable to assume that an uninformed principal would choose a mediator that is ex ante incentive efficient. The P-Max Mediator maximizes the principal's ex ante expected payoff among the IIE mediators. Thus, for uninformed mediator selection, *the P-Max Mediator is the only reasonable choice by the uninformed principal* while any other IIE mediators are not plausible choices.[13]

In informed mediator selection, because the principal already knows her type when choosing, she would be concerned with her interim expected payoff in mediation given her true type. A naive intuition might suggest that the principal would choose a mediator among the IIE mediators to maximize her interim expected payoff given her true type. In our setting, the Neutral Mediator is the best feasible mediator for the H type, whereas the P-Max Mediator is the best feasible mediator for the L type. If the principal is expected to choose the feasible mediator that her type most prefers, then the subordinate would be able to infer the principal's type based on her mediator choice. And the chosen mediator would become infeasible (either not incentive compatible or not individually rational) as soon as it is selected. Then, how and which mediator should the informed principal choose?

In our environment, any IIE mediator that is selected with probability one regardless of the principal's type (and the players participate and honestly report their types in mediation thereafter) can be supported as a sequential equilibrium of the informed mediator selection game.[14] Moreover, there is no separating equilibrium in which different types choose distinct mediators followed by the players' honest participation.[15] This gives us a behavioral prediction that *two principal types would pool on the same IIE mediator*.[16]

[13] We use the weak implementation concept. That is, given the requirement of incentive efficiency at the ex ante stage, the uninformed mediator selection game has an equilibrium supporting the P-Max Mediator but other IIE mediators cannot be sustained in equilibrium.

[14] See Myerson (1983, Sec. 5) and Kim (2017, fn.19).

[15] Any separation, where two principal types choose different IIE mediators (including randomization), cannot be part of an equilibrium, because the fact that a particular mediator is chosen allows the subordinate to learn about the principal's type. See Appendix B.5 for a formal proof.

[16] This result is related to but not a direct consequence of applying the inscrutability principle (Myerson, 1983), according to which there is no loss of generality in assuming that all types of the principal should choose the same mediator. This principle is an elegant analytical tool that enables one to focus on the set of pooling equilibria wherein all principal types choose the same mediator for fully characterizing the set of equilibrium outcomes. However, it does not mean to generate any specific behavioral prediction.

However, the concept of sequential equilibrium, as well as standard refinements used in the literature, cannot determine a narrower prediction about which IIE mediators are more likely to be the pooling outcomes. Going beyond the existing noncooperative solution concepts, Myerson (1983) develops a theory of inscrutable mechanism selection by an informed principal; and provides several notions of how the informed principal should make an *inscrutable intertype compromise* (Myerson, 1991, p.523): To be inscrutable, the predicted mediator must reflect some kind of compromise between the different goals of the principal's different types. That is, when the principal's possible types prefer different IIE mediators, the principal must choose the one that will be perceived as a reasonable compromise between what she really wants and what she might have wanted if her type had been different. In fact, our pooling equilibrium outcomes are reminiscent of such an idea in the weak sense of satisfying sequential rationality in the honest participation equilibrium, predicting that any IIE mediator can be perceived as an inscrutable intertype compromise. The concept of neutral optimum (Myerson, 1983) further refines the notion of inscrutable intertype compromise, predicting that among the IIE mediators *the Neutral Mediator is the most reasonable choice by the informed principal.*

The intuition can be explained as follows. On the one hand, expressing a preference for the P-Max Mediator would convey information that the principal is an L type. An H type subordinate will then be convinced to immediately go to war when matched with such a principal. On the other hand, expressing a preference for the Neutral Mediator would convey information that the principal is an H type, in which case an L type subordinate will be convinced to lie in implementing the mediator when matched with such a principal. But if the principal is the H type, then she could argue for hiring the Neutral Mediator and add the following statement: "If you infer from my preference for the Neutral Mediator that my type is H, then you should dispense with both mediators and we can just split the surplus in a way that would be better for you and just as good for me when I am the H type." The L type cannot make the same argument in favor of the P-Max Mediator.[17] In a sense, the H type would actually be very eager to reveal its type, whereas the L type would have an

[17] This is because there is no way for the L type principal to propose a split of the surplus in a way that would be better for the H type opponent than what he can achieve, having inferred that the principal's type is L, and just as good for the principal herself as in implementing the chosen mediator.

incentive to conceal its type. The L type principal, in particular, must make this intertype compromise and so would have to mimic the H type by doing what the H type would do, to maintain inscrutability. Thus, the informed principal would choose the mediator that is most preferred by the H type, never revealing her type during the selection process.[18]

Importantly, the key component of the inscrutable intertype compromise is that the principal must contemplate interpersonal comparisons between the two possible types (although she already knows what her type is) through an *implicit* or virtual thought process *before* actually choosing a mediator.

3 Experimental Paramaterization and Hypotheses

We conduct two experiments: Experiment I in which either informed or uninformed subjects choose a mediator, and Experiment II in which informed subjects choose a scheme of selecting a mediator for their own type and the unrealized type. The first experiment allows us to test the principal's mechanism selection problem as well as to study how subjects respond to possessing private information at the time of selecting a mediator. The second experiment is carried out for the purpose of scrutinizing informed subjects' behavior of choosing a mediator. In this section, we provide experimental paramaterization and hypotheses that govern both experiments. Sections 4 and 5 will present more details about the experimental procedures for Experiments I and II, respectvely.

3.1 Model Parameters and Two Mediators

We fix $\theta = 0.75$ and $\delta = 0.8$ throughout the experiments. These parameters are chosen so that the two types' different preferences over outcomes are salient. In Experiment I, we consider two different values for the prior probability of H type: $q = 1/4$ and $q = 2/5$. In Experiment II, we focus on the case of $q = 1/4$. For each q, two mediators that correspond to the P-Max and Neutral Mediators are available to experimental subjects as possible mediator choices. The restriction to those two choices does not change the theoretical predictions

[18] This assertion holds for any probability of H type under which the two types prefer different IIE mediators. In a sense, the H type is more influential on the players' behavior in choosing a mediator even if its proportion is very small.

established in Section 2.3, and simplifies the subjects' problem without complicating the data with random choices.[19]

The mediation plans of the two mediators that we use in the lab are shown in Table 1. In the lab, we phrase each mediator according to its associated value of p_M (as specified in brackets in the table) instead of its theoretical name. The table also shows the ex ante probability of agreement under each mediator, denoted by $P(peace)$.

Table 1: Two Mediators

		p_L	p_M	p_H	$P(peace)$
$q = 1/4$	P-Max* [40-Mediator]	1	.40 ($\frac{5}{12} \approx .4167$)	1	77.5% (78.1%)
	Neutral [0-Mediator]	1	0	1	62.5%
$q = 2/5$	P-Max* [70-Mediator]	1	.70 ($\frac{75}{103} \approx .7282$)	.85 ($\frac{630}{721} \approx .8738$)	83.2% (84.9%)
	Neutral* [0-Mediator]	1	0	.50 ($\frac{15}{28} \approx .5357$)	44% (44.6%)

Note: The asterisks indicate that the mediators' probabilities are approximations of the theoretical P-Max and Neutral Mediators, the exact values for which are shown in parentheses.

In using approximations, we had two objectives. First, we chose to make the descriptions of mediators simple enough for subjects to understand. Using integer values (in percentage) would help subjects easily understand and compare mediation plans. Second, we chose to avoid subjects' indifferences due to binding incentive constraints in order to help making incentives salient in the experiment. Table 2 shows whether the L and H types' incentive compatibility (IC) and individual rationality (IR) constraints bind or not under the theoretical P-Max and Neutral Mediators. With a binding H-IR constraint, an H type would

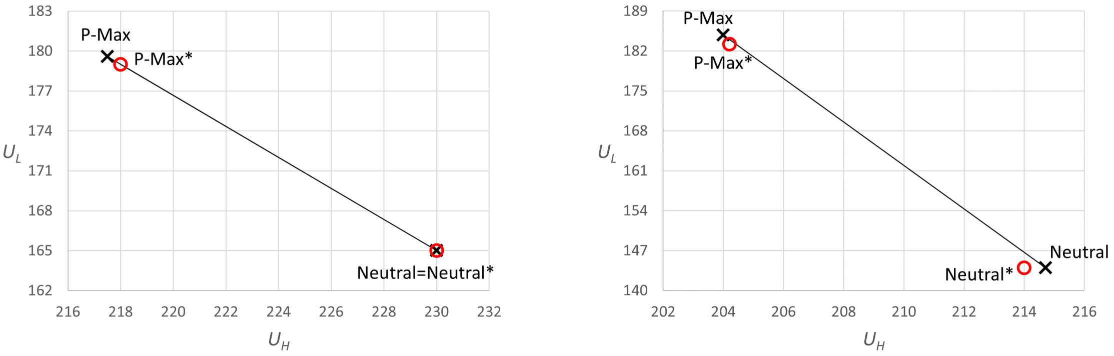
Table 2: IC and IR Constraints in Theory

		H-IR	L-IR	H-IC	L-IC
$q = 1/4$	P-Max ($p_M = 5/12$)	bind	not bind	not bind	not bind
	Neutral ($p_M = 0$)	not bind	not bind	not bind	not bind
$q = 2/5$	P-Max ($p_M = 75/103$)	bind	not bind	not bind	bind
	Neutral ($p_M = 0$)	not bind	not bind	not bind	bind

[19] We believe that adding more IIE mediators (or providing the entire set of IIE mediators) will not change the qualitative insights of the experimental results. See Appendix B.4 for relevant discussions.

be indifferent between participating in mediation and choosing disagreement; and with a binding L-IC constraint, an L type would be indifferent between sending two messages l and h. By using $p_M = .40$ and $.70$ for the P-Max Mediators, the H-IR constraints become slack. By using $p_H = .85$ and $.50$ for the P-Max and Neutral Mediators, respectively, when $q = 2/5$, the L-IC constraints become slack.[20]

Figure 1 shows the IIE frontier, which is the set of IIE payoff pairs of H and L types, (U_H, U_L), under the theoretical IIE mediators; as well as the expected payoff pairs under the two mediators that we use in the lab. With all the incentive constraints slack, P-Max* Mediator for $q = 1/4$ is still on the IIE frontier, while P-Max* and Neutral* Mediators for $q = 2/5$ fall short of but sufficiently close to the frontier. Because we are interested in subjects' choice between the two extremes along the IIE frontier and not in their choice to achieve the IIE frontier per se, we consider the two mediators used in our experiments to be reasonable approximations to the theoretical counterparts.



(a) For $q = 1/4$

(b) For $q = 2/5$

Note: The line depicts the IIE frontier of payoff pairs under the IIE mediators. The $\times$ markers indicate the expected payoffs under the theoretical P-Max and Neutral Mediators. The circle markers indicate the expected payoffs under the approximations, P-Max* and Neutral* Mediators.

Figure 1: The Expected Payoffs of H type (U_H) and L Type (U_L) in IIE Mediators

The reason for considering two values of q is as follows. When $q = 1/4$, the two mediators differ only in terms of the values of p_M. Their simplicity may help subjects easily compare

[20] When $q = 2/5$, p_H must satisfy $p_H \leq 241/280 \approx .8607$ given $p_M = .7$, and $p_H \leq 15/28 \approx .5357$ given $p_M = 0$. These two inequalities characterize the L-IC constraints given p_M. Setting p_H as high as possible maximizes the H type's interim expected payoff without affecting the L type's interim expected payoff. Thus there is a trade-off between making the L-IC constraint slack and achieving interim incentive efficiency.

the two mediators when choosing. However, knowing that $q = 1/4$, subjects might neglect the possibility of H type being realized. When $q = 2/5$, such a possibility is greater but the two mediators look more complex than when $q = 1/4$. Collecting data under both $q = 1/4$ and $q = 2/5$ allows us to see whether the probability of type or the complexity of mediation plans affects subject behavior.

As can be seen in Figure 1, the difference in the L type's expected utility payoffs between the two mediators when $q = 2/5$ is much larger compared to that when $q = 1/4$, while that difference for the H type is slightly smaller when $q = 2/5$ relative to when $q = 1/4$. Also, $P(peace)$ between the two mediators are more apart when $q = 2/5$ than when $q = 1/4$. We do not expect that subjects would exactly calculate all those measures during the experiment. Importantly, for both $q = 1/4$ and $q = 2/5$, the underlying payoff structure (i.e., how the actual payoffs are allocated depending on their types under agreement and disagreement) is the same and so all the relevant incentives of different types of players are salient.

3.2 Hypotheses

Table 3 summarizes the theoretical predictions for mediator selection by informed and uninformed principals when given the P-Max and Neutral Mediators (Section 2.3).

Table 3: Theoretical Predictions for Principal's Mediator Selection

Info structure	Uninformed	Informed	
Requirement	Ex Ante Incentive Efficiency	Interim Incentive Efficiency	Neutral optimum
Selection in Equilibrium	P-Max Mediator	Compromising on either P-Max or Neutral	Compromising on Neutral Mediator

Based on the theoretical predictions, we present three testable hypotheses. The first hypothesis regards the theory of efficient mechanisms for the principal's mechanism selection before any player has private information.

Hypothesis 1 (Uninformed Mediator Selection). *In the case of uninformed mediator selection, the uninformed principal chooses the P-Max Mediator over the Neutral Mediator.*

Transitioning to the informed mediator selection environment, our second hypothesis aims to explore how informed principals navigate inscrutable intertype compromises under the conflicting objectives between their true type and their alternate potential type. We intend to assess these compromises through two distinct approaches. In Experiment I, we will analyze the choice data to determine if both principal types opt for the same mediator. Consistent observations would immediately lend support to the notion of inscrutable intertype compromise. However, divergent behavior should not immediately be taken as evidence against the inscrutable intertype compromise but would prompt further investigation into the underlying reason that leads to such a behavior. This is because observed divergent behavior may arise either from a lack of comprehension regarding the value of inscrutable choice or from a discrepancy in their inscrutable intertype compromise between the two types, each "compromising" on different mediators. In Experiment II, we will explore the informed principals' perspectives on what mediator their unrealized type would take in formulating their intertype compromise. This will involve eliciting the subject's mediator selection based on their informed type, as well as their hypothetical choice if they were to embody the alternative type. Each subject should "phrase" her mediator choice in a way that is independent of her type because any type-revealing choice of mediator would cease to be incentive feasible.

Hypothesis 2 (Informed Mediator Selection - Inscrutable Intertype Compromise)**.** *In the case of informed mediator selection,*

(a) In Experiment I, both H and L types of informed principals choose the same mediator.

(b) In Experiment II, each type of informed principal believes that her other unrealized type should choose the same mediator as herself.

Our next hypothesis tests Myerson's theory of neutral optimum. While both the P-Max and Neutral Mediators correspond to the sequential equilibrium predictions of the mechanism selection game, the concept of neutral optimum makes a prediction toward the Neutral Mediator among the two. Thus rejecting Hypothesis 3 would imply that the theory of neutral optimum may not work in an experimental setting.

Hypothesis 3 (Informed Mediator Selection - Neutral Optimum)**.** *In the case of informed mediator selection, the informed principal chooses the Neutral Mediator over the P-Max Mediator.*

We have three additional questions to ask in the informed mediator selection environment:

(1) Do H type principals choose the Neutral Mediator more often than do L type principals?

(2) What inferences do subordinates make conditional on either the P-Max or Neutral Mediator chosen by the principal?

(3) Is the P-Max Mediator (or the Neutral Mediator), when chosen by the informed principal, declined by subordinates? If so, by which type of subordinates?

If H type subjects tend to choose the Neutral Mediator while L type subjects tend to choose the P-Max Mediator, then that would be suggestive evidence of subjects' understanding of their type's preferred mediator. The second and third questions above relate to the driving force behind the theory of neutral optimum: the P-Max Mediator is not expected to be selected by the players precisely because H type subordinates would choose to decline rather than let it be selected if they update their beliefs that the principal is more likely to be an L type after observing the principal's choice of P-Max Mediator.[21] Examining observations from the data on subjects' inferences and strategies of type-reporting and declining would enable us to explain why Myerson's theories for informed principal problems work or not work in the lab.

4 Experiment I

4.1 Experimental Treatments and Procedure

The treatment variables are the prior probability of types ($q = 1/4$ or $2/5$) and the information structure at the time of mediator selection (uninformed or informed). Table 4 summarizes our 2×2 treatment design. Each of those treatments will be referred to as Simple-Uninformed ($q = 1/4 +$ uninformed selection), Simple-Informed ($q = 1/4 +$ informed selection), Complex-Uninformed ($q = 2/5 +$ uninformed selection), and Complex-Informed ($q = 2/5 +$ informed selection). The treatments under $q = 2/5$ are labeled "Complex" because the given mediators have different values of p_H that correspond to two different values

[21] Yet in theory subordinates should maintain holding their prior beliefs about the principal's type regardless of the principal's chosen mediator if they perceive the mediator choice to be non-type-revealing.

of p_M, whereas the mediators under "Simple" treatments only differ in terms of p_M. We ran 4 sessions for each of the four treatments.

Table 4: Experimental Treatments

		Probability of H type	
		$q = 1/4$	$q = 2/5$
Info structure	Uninformed	Simple-Uninformed	Complex-Uninformed
	Informed	Simple-Informed	Complex-Informed

We implemented the experimental design in which each session had 4 separate parts, labeled in the paper as UC, M1, M2, U-MS or I-MS. Each of the UC, M1, and M2 parts consisted of 4 rounds, and the U-MS (for Uninformed treatment) or I-MS (for Informed treatment) consisted of 28 rounds. We always ordered UC first and U-MS/I-MS last. The UC part corresponds to the unmediated communication game described in Appendix C.1. In the M1 and M2 parts, subjects play the mediation game described in Section 2.1, with each part given one of the two mediators. Across the four sessions for each treatment, we varied the order of M1 and M2 so as to treat the two mediators symmetrically. The last part implements the main treatment of either uninformed mediator selection (U-MS) or informed mediator selection (I-MS), in which subjects play the corresponding mediator selection game described in Section 2.2. The three preceding parts are akin to practice rounds, providing the subjects with some understanding of the underlying situation and how mediation works with different available mediators. The details of the experimental procedures for Parts 1-3 are provided in *Experimental Instructions*. We will now describe the procedure for Part 4, where the key observations needed to test our hypotheses are generated.

Informed Mediator Selection (I-MS). The I-MS part was presented as the last part of eight experimental sessions, four under Simple-Informed and another four under Complex-Informed. At the start of each round, subjects were matched in pairs, randomly and anonymously, and independently assigned types by the computer according to q. After learning their own type, subjects were asked to choose which mediator to rely on in the mediation among the two mediators, imagining themselves as being the selector of a mediator.[22] One of the choices made by two subjects in a pair was randomly, with equal chances, chosen.

[22] In the experiment, we used the terms "selector" and "non-selector" instead of principal and subordinate.

Then the selector of the mediator for the round and which mediator the selector has chosen were announced. The non-selector was asked to make some inferences about the selector's type based on her chosen mediator.[23] Next, both subjects each sent to the chosen mediator a confidential message among $\{H, L\}$ but the non-selector had an additional option to decline the mediator instead of sending a message. If the non-selector chose to decline, then disagreement occured immediately and each subject received payoffs according to the subjects' true types. Otherwise, given reported types, the mediator prescribed an agreement or walked out, and the payoffs were realized. At the end of each round, subjects received feedback on the types, mediator choices by two subjects, the selector and the selector's mediator choice, whether the non-selector declined, the messages sent to the mediator if played, the final outcome, and one's own payoff.

Uninformed Mediator Selection (U-MS). The U-MS part was presented as the last part of eight experimental sessions, four under Simple-Uninformed and another four under Complex-Uninformed. At the start of each round, subjects were matched in pairs, randomly and anonymously. As in the I-MS, subjects were asked to choose which mediator to rely on among the two mediators, imagining themselves as being the selector of a mediator, but without knowing their own type or the partner's type. Subjects only knew that their types were each likely to be H or L according to q. One of the choices by two subjects in a pair was randomly, with equal chances, chosen. The selector of the mediator for the round and which mediator the selector has chosen were announced, and then subjects were informed of their assigned private types. The subsequent stages of inference and implementation, as well as the realization of the outcome followed the same procedures as in the I-MS part.

We conducted the experiment in English using oTree in real-time online mode via Zoom at the Hong Kong University of Science and Technology (HKUST). A total of 298 subjects were recruited from the graduate and undergraduate population of the university. Each

[23] Our primary objective was to give subordinate (non-selector) subjects an opportunity to infer and update their beliefs based on the principal's mediator choice. To facilitate this, we provided participants with three options for reporting their posterior beliefs: (i) More likely to be H (than the prior), (ii) Same as the prior, and (iii) More likely to be L (than the prior). Examining how subjects update their beliefs can be useful to understand the observed behavior, but it is not part of the description of the game. So, we did not incentivize the belief elicitation. Thus, the data obtained from the belief elicitation was treated as supplementary.

subject participated in one of the 16 ($= 4 \times 4$) sessions. Session sizes varied from 16 to 20. In each session, upon arrival at the designated Zoom meeting, subjects were instructed to turn on their videos during the entire course of the experiment. Each session lasted 1.5 hours on average. To avoid decimals, the size of the surplus in the game was set to 400 points. The number of points that each subject earned at one randomly selected round was converted into HKD at the rate of 1 point=1 HKD. The payments ranged from HKD 100 to 280 including the HKD 40 show-up fee, with an average of HKD 220 ($\approx$ USD 28.20). The experimental instructions for Simple-Informed/Uninformed treatments are provided in *Experimental Instructions*.

4.2 Experimental Results

We report our experimental results obtained by aggregating data across all rounds in each part of all sessions within the same treatment. It is important to note that the qualitative findings remain consistent regardless of whether we consider all rounds or a selected subset for each part. Table 5 in Appendix A provides the non-parametric test results for further reference.[24] The results exhibit a high degree of consistency across the Simple and Complex environments. Therefore, unless specifically stated otherwise, we will describe our results collectively without distinguishing between the Simple and Complex treatments.

4.2.1 Principal's Mediator Choice

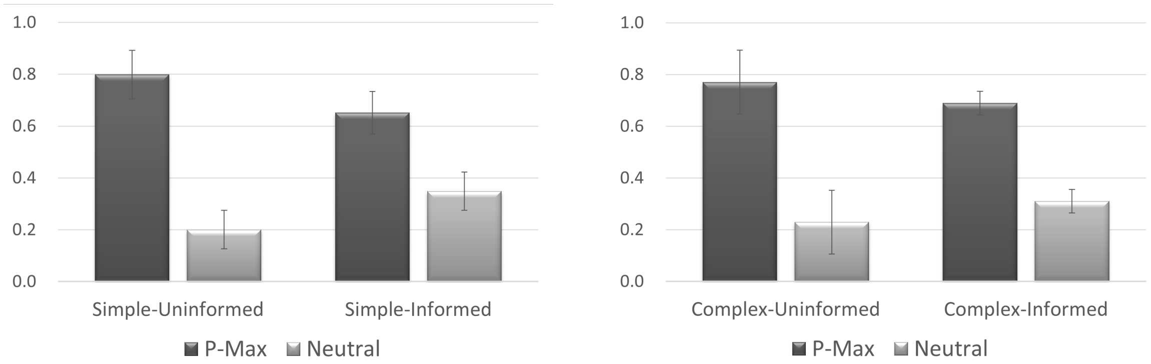
We begin by analyzing the data from the main part of our experiment (Part 4) in which the subjects made their mediator choice. The subjects were unaware of their randomly assigned role as a principal or a subordinate when making their mediator choice. So examining the data from all subjects rather than only principal-subjects is appropriate.

The two panels of Figure 2 display the proportions of two mediators chosen by all subjects with their types pooled in each treatment for all four treatments.[25] In each panel, the

[24] When describing the results in this paper, "marginally significant," "significant," and "insignificant" refer to the case in which the p-value from the non-parametric test is between 0.05 and 0.1, strictly below 0.05, and strictly above 0.1, respectively. We add a caveat that with four independent observations (sessions) in Experiment I, the minimum attainable p-value is 0.0625 (one-sided) for the Wilcoxon signed-rank test, in which case we will simply refer to the case as being "predominant."

[25] Figure 16 in Appendix D.1 shows the proportions of mediators chosen by principal-subjects and played

Uninformed treatment data are reported on the left, and the Informed treatment data on the right.[26]



(a) Simple

(b) Complex

Figure 2: Proportion of Mediator Chosen (All Subjects)

It is immediately evident that the subjects in the Uninformed treatments choose the P-Max Mediator predominantly more frequently than the Neutral Mediator, consistent with the theoretical prediction for the uninformed principal's mediator selection problem. In particular, 79.9% of subjects choose the P-Max Mediator while 20.1% of subjects choose the Neutral Mediator in Simple-Uninformed; 77.1% of subjects choose the P-Max Mediator while 22.9% choose the Neutral Mediator in Complex-Uninformed. These observations lead to the following finding.

Finding 1. *Consistent with Hypothesis 1, in the case of uninformed mediator selection, the uninformed principal chooses the P-Max Mediator more often than the Neutral Mediator.*

For the case of informed mediator selection, as can be seen in the Informed treatment data on the right in each panel of Figure 2, the subjects do not choose the Neutral Mediator over the P-Max Mediator; rather they choose the P-Max Mediator predominantly more often (65.2% in Simple; 68.9% in Complex) than the Neutral Mediator (34.8% in Simple; 31.1% in Complex).[27] This result indicates clear evidence that the prediction by the theory of neutral optimum does not hold in our experimental setting. We summarize this result as follows:

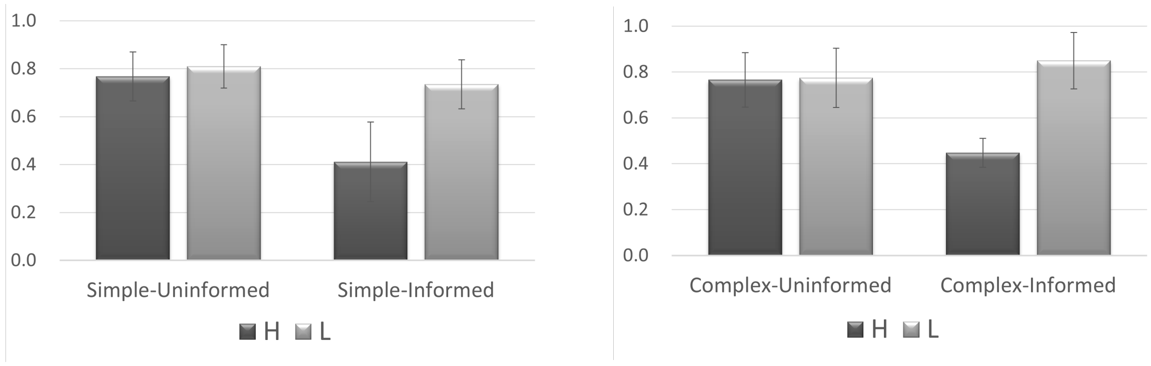
after rejections for the four treatments. The pattern of the graph is essentially the same as in Figure 2.

[26] In every bar graph presented in this paper, we show 95% confidence intervals calculated from standard errors clustered at the session level.

[27] The informed subjects choose the P-Max Mediator over the Neutral Mediator significantly less frequently than do the uninformed subjects (one-sided p-value<0.05, Mann-Whitney test).

Finding 2. *Counter to Hypotheses 3, in the case of informed mediator selection, the informed principal chooses the P-Max Mediator more often than the Neutral Mediator.*

Figure 3 shows the proportions of the H and L types of all subjects that choose the P-Max Mediator for each treatment.[28] Note that the complementary probability for each bar is the proportion of the respective type subjects that choose the Neutral mediator. The two bars on the right in each panel of Figure 3 show that the proportion of L type subjects who choose the P-Max Mediator is predominantly higher (74% in Simple; 85% in Complex) than the proportion of H type subjects who choose the P-Max Mediator (41% in Simple; 45% in Complex). Thus, we have the following result:



(a) Simple

(b) Complex

Figure 3: Proportion of P-Max Mediator Chosen By Type (All Subjects)

Finding 3. *Counter to Hypotheses 2(a), in the case of informed mediator selection, the two principal types do not choose the same mediator.*

Figure 3 also shows that in informed mediator selection, the L type subjects choose the P-Max Mediator predominantly more often than the Neutral Mediator, while the H type subjects choose the Neutral Mediator more often than the P-Max mediator. This behavior was consistent over all rounds. We take this as evidence that the majority of subjects appear to recognize that, given their prior beliefs, the P-Max Mediator is preferred by the L type

[28] Figure 17 in Appendix D.1 shows the proportions of the two types of principals that choose the P-Max Mediator over the Neutral Mediator for each treatment. Behavior of all subjects is essentially the same as that of principal-subjects.

while the Neutral Mediator is preferred by the H type.[29]

Note that we cannot immediately conclude that the observations in Figure 3 contradict Hypothesis 2(b). The reason for different types choosing different mediators might not be that subjects are not enacting the inscrutable intertype compromise, but rather might be that different types focus on different inscrutable intertype compromises. We examine this issue in Section 5 where we test whether and how subjects consider implicit compromise on mediator choice.

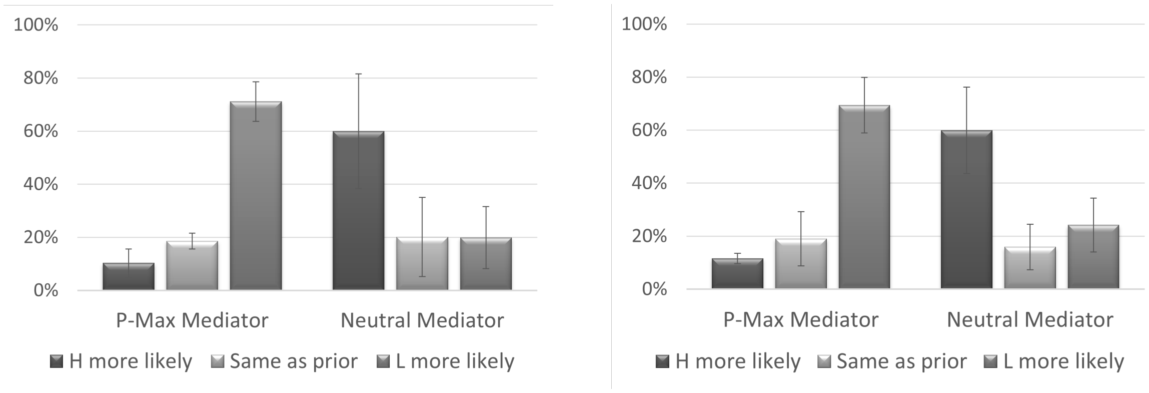
4.2.2 Subordinate's Inference and Strategy

We provide observations from our data on subordinate subjects' inference and strategy during the mediation subgame. Examining those data gives some indications of why the concept of neutral optimum was not realized in the lab.

For experimental subjects to make the inscrutable intertype compromise in the way that the concept of neutral optimum predicts, the following aspects of their understanding must be in check. First, subjects must understand that information can be revealed by their mediator choice if they choose their preferred mediator. Second, subjects should realize that, only L type principals would have some incentive to conceal their type because H type subordinates can benefit from declining the L type-revealing choice of P-Max mediator. Importantly, subjects must speculate on these considerations *before* selecting a mediator, realizing the benefit of inscrutably selecting the Neutral Mediator, and be induced to choose the Neutral Mediator in the first stage.

After the principal's chosen mediator is announced, a subordinate chooses one of the three inferences about the principal's type: (i) More likely to be H (than the prior), (ii) Same as the prior, and (iii) More likely to be L (than the prior). Figure 4 reports the frequencies of three inferences by subordinates conditional on either the P-Max Mediator or the Neutral Mediator chosen by the principal. As can be seen in both panels of Figure 4, subordinates infer that the principal is more likely to be the L type after observing the principal's choice of the

[29] The two bars on the left in each panel of Figure 3 show that the proportions of both types choosing the P-Max Mediator are high (at almost 80%) in the Uninformed treatments. This has no substantive implication for subject behavior by type because the subjects choose a mediator before knowing their types in the Uninformed treatments, but simply implies that the subjects understand that the P-Max Mediator is better ex-ante for both types in uninformed mediator selection.



(a) Simple-Informed

(b) Complex-Informed

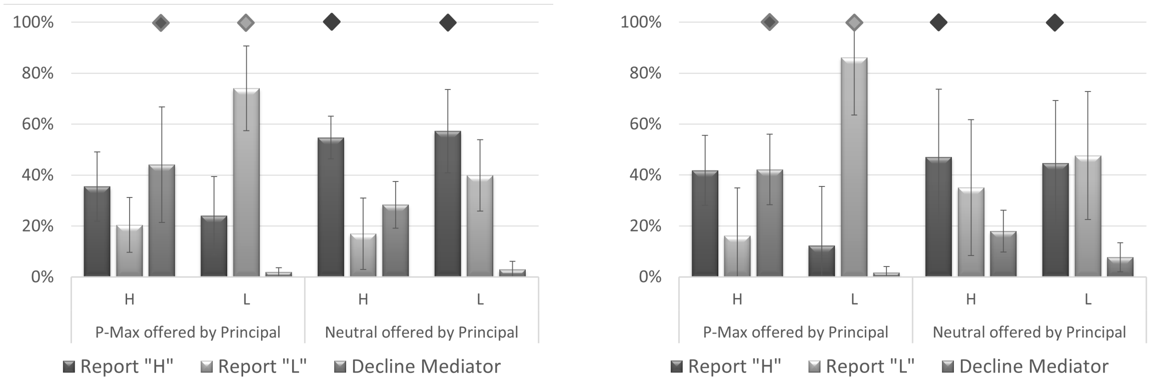
Figure 4: Subordinate's Inference Conditional on Each Mediator Chosen (Informed)

P-Max Mediator with predominantly high frequency (about 70%); and when the principal has chosen the Neutral Mediator, the proportions of subordinates updating their beliefs that the principal is more likely to be the H type are predominantly high (about 60%). Also, the frequencies of three inferences reported in Figure 4 were remarkably consistent over all rounds. These are suggestive evidence that the subjects would understand that information is revealed by their mediator choice.

Figure 5 shows the frequencies of different strategies that subordinates choose (by type) conditional on either the P-Max Mediator or the Neutral Mediator chosen by the principal for the Informed treatments.[30,31] We observe that the frequencies that the H type subordinates decline the P-Max Mediator chosen by the principal are 44% in Simple-Informed and 42% in Complex-Informed. This is of particular interest because for the theory of neutral optimum to work in our setting, L type subjects must consider that choosing the P-Max Mediator may reveal their type and thus be declined by H type subordinates; and H type subordinates must consider that the choice of P-Max Mediator is made by an L type and thus decline it. These considerations could be learned through 28 rounds of mediator selection *if* the choice of P-Max Mediator actually led to more rejections by H type subordinates. However, in the data, not only the rejection frequencies are low for L type principals to experience the

[30] Figures 19 and 20 in Appendix D.1 show the frequencies of different messages that principals send (by type) conditional on either P-Max or Neutral Mediator chosen for all four treatments.

[31] Figures 21–24 in Appendix D.1 show the frequencies of different strategies that subordinates choose (by type) given each possible inference conditional on either P-Max or Neutral Mediator chosen by the principal for all four treatments.



(a) Simple-Informed

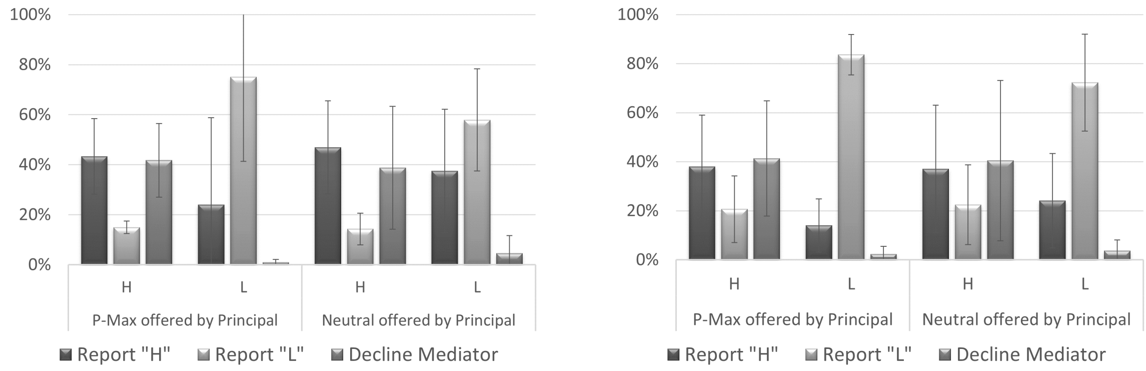
(b) Complex-Informed

Note: The diamond shapes indicate the subordinate's best response of each type to the *predominant* reported belief about the principal's type given the chosen mediator (that L more likely than the prior given P-Max and H more likely than the prior given Neutral, shown in Figure 4 and confirmed by the Wilcoxon signed-rank test). Note that the best responses are computed against the principal reporting truthfully in implementing her chosen mediator (see Appendix C.2).

Figure 5: Subordinate's Strategy in Chosen Mediator by Type (Informed)

consequence of choosing the P-Max Mediator but also those remain consistently low over rounds.

Another interesting observation in Figure 5 is that there is some tendency for H type subordinates to decline even their preferred Neutral Mediator when offered by the principal, although the rate is not so high (28% in Simple-Informed and 18% in Complex-Informed).[32] That tendency is more stark in the Uninformed treatments, as can be seen in Figure 6. The



(a) Simple-Uninformed

(b) Complex-Uninformed

Figure 6: Subordinate's Strategy in Chosen Mediator by Type (Uninformed)

[32] The L type subordinates decline only 1.6–7.7% of the times across all four cases in Figure 5.

frequencies that H type subordinates decline the chosen mediator range consistently from 39% to 42% across the four cases. Also, the difference in the frequency of H type declining the chosen mediator (either P-Max or Neutral) between the Uninformed and Informed treatments is statistically insignificant.

We summarize the findings that answer the three questions posed in Section 3.2.

Finding 4. *In the case of informed mediator selection,*

(1) The Neutral Mediator is chosen over the P-Max Mediator more often by H type principals than by L types.

(2) Given the principal's choice of P-Max (resp., Neutral) Mediator, the subordinates infer that the principal is more likely to be an L type (resp., H type).

(3) L type subordinates hardly decline the principal's chosen mediator. The P-Max Mediator, when chosen by the informed principal, is declined by H type subordinates 42-44% of the time on average; The Neutral Mediator is also declined although the rate is only about 28% or lower. A similar and more consistent pattern of rejection (39-42%) is observed in the case of uninformed mediator selection.

Comparing Figures 5 and 6, we observe some similar patterns of subordinates' strategies in both Informed and Uninformed treatments, except for the L types being more sincere to the Neutral Mediator in the Uninformed treatment than in the Informed treatment. However, the inference data are in stark contrast between Informed and Uninformed (cf. Figure 4 and Figure 18 in Appendix D.1). In particular, after observing either the P-Max or the Neutral Mediator chosen, the frequency of the "same as the prior" inference is significantly higher in Uninformed treatment than in Informed treatment (Mann-Whitney test, one-sided p-values<0.05). This means that although the subjects seem to understand that no information can be inferred in uninformed mediator selection unlike in informed mediator selection, they tend to behave similarly when actually playing the chosen mediator in both cases.

In the lab, each type of subjects tends to choose her preferred mediator, information is indeed revealed by the mediator choice, and the subjects make corresponding inferences about the principal's type. But the subjects' inferences appear to have little effect on their subsequent plays in implementing the mediator, as well as on what mediator they would initially choose even through multiple rounds. As a result, the inscrutable intertype compromise that should be resolved in favor of the H type in the theory of neutral optimum does not occur in the lab. Our experimental evidence suggest the driving forces behind this failure to be that H types do not fully act on their inferences and that L types do not fully anticipate what their choice may bring. But this does not imply that the subjects do not contemplate inscrutable intertype compromise when deliberating over their mediator selection. To see how the inscrutable intertype compromise gets resolved and to identify the primary source of the failure of the theory of neutral optimum, we conducted the second experiment.

5 Experiment II

5.1 Inscrutable Intertype Compromise

Figure 3 shows that the two types of subjects do not choose the same mediator; in particular, L types choose the P-Max Mediator more often while H types choose the Neutral Mediator more often. This empirically-observed type-dependent choice may stem from subjects not understanding the benefit of inscrutable mediator selection, in which case subjects would be honing in on nonequilibrium play of separation. Alternatively, it could occur when subjects recognize the benefit of inscrutable mediator selection and choose inscrutably but their two types struggle to reach an agreement on their intertype compromise.

To identify which scenario is relevant and to better understand participant behavior, we conducted the second experiment. Initially, the principals were informed of their types. We then offered them three options that specify not only their informed-type's choice of mediator but also what their choice would have been if they belonged to the other type:

- Neutral Compromising: choose the Neutral Mediator and select the same one if they were of the other type.
- P-Max Compromising: choose the P-Max Mediator and select the same one if they were of the other type.

- Uncompromising: choose the Neutral Mediator if the principal is of the H type and the P-Max Mediator if the principal is of the L type.[33]

The main objective of this design is to understand what each informed principal, who has a clear objective of maximizing her interim utility, thinks her other unrealized type should do in making the intertype compromise.[34] We test whether the informed principals choose inscrutably and how they resolve their inscrutable intertype compromise (Hypothesis 2(b)).

5.2 Experimental Treatment and Procedure

Recall that we fix $\theta = 0.75$ and $\delta = 0.8$ for the baseline model. For this experiment, we focus on the Simple-Informed case with $q = 1/4$ where the two mediators are 40-Mediator (P-Max) and 0-Mediator (Neutral).

We implemented the experimental design in which each session had 4 separate parts, labeled in the paper as R1, R2, R3, I-RS. Each of the R1–R3 parts consisted of 8 rounds, which let subjects experience each of the three options (Neutral Compromising, P-Max Compromising, and Uncompromising), which are phrased in the lab as mediator-selection rules without reference to their theoretical names. The last part I-RS consisted of 16 rounds, in which the informed principal selects one of the three mediator-selection rules. In this last part, subjects are essentially playing the informed principal's mediator selection game by choosing a mediator-selection rule that assigns their chosen mediator. The details of the experimental procedures for Parts R1-R3 are provided in *Experimental Instructions*. We will now describe the procedure for Part I-RS, where the key observations needed to test our hypotheses are generated.

[33] Another possible uncompromising option is to choose Neutral if L type and P-Max if H type; but such possibility is clearly unreasonable because Neutral is H type preferred and P-Max is L type preferred, so we do not consider such an option.

[34] A conventional design employing the strategy method (Selten, 1967) will not enable us to achieve this objective because it only allows each informed principal to think about her choice based on her given type (by asking the principal to specify her type-contingent plan), but not for her other unrealized types. The intertype compromise requires a thought process in the mind of each type of informed principal regarding the behavior of the unrealized types. Furthermore, in a Bayesian game where the sender's strategy involves a type-contingent message plan, the choice method implemented with standard experimental procedures–where types are randomly drawn in each round–is essentially equivalent to the strategy method.

Informed Rule Selection (I-RS). At the start of each round, subjects were matched in pairs, randomly and anonymously. One subject in a pair was randomly, with equal chances, chosen and announced to be the selector of a mediator. Subjects were independently assigned private types by the computer according to $q = 1/4$. After learning one's own type, the selector of a mediator was asked to choose which mediator-selection rule, among the three options, to be used to select a mediator. Then the selector's mediator choice was automatically assigned based on the mediator-selection rule that the selector had chosen and the selector's type. The selector's chosen mediator was announced; at this stage, the non-selector did not know which mediator-selection rule the selector had chosen but only observed the selector's mediator choice. The rest of the negotiation process is the same as those in previous parts. At the end of each round, subjects received feedback on the selector's chosen mediator-selection rule, the types, the selector's mediator choice (according to the mediator-selection rule), whether the non-selector declines, the messages sent to the mediator if played, the final outcome, and one's own payoff.

We conducted the experiment in English using oTree in real-time online mode via Zoom at HKUST. A total of 90 subjects were recruited. We ran 6 sessions in one of which each subject participated. Session sizes varied from 6 to 22. Each session lasted 1.5 hours on average. The number of points that each subject earned at one randomly selected round was converted into HKD at the rate of 1 point=1 HKD. The payments ranged from HKD 100 and 280 and the average payment was HKD 225.7 ($\approx$ USD 29). The experimental instructions for this experiment are presented in *Experimental Instructions*.

5.3 Experimental Results

We report our experimental results from the second experiment, obtained by aggregating data across all rounds in each part of all sessions. Table 6 in Appendix A provides the relevant non-parametric test results for further reference.

5.3.1 Principal's Inscrutable Intertype Compromise

We analyze the data from Part 4 in which principals make their choice of mediator-selection rule. Figure 7 displays the proportions of three mediator-selection rules chosen by all principals with their types pooled. We observe that the Uncompromising rule is rarely chosen (22.5%) relative to the two Compromising rules together (77.5%). This observation indicates that the majority of informed principals compromise with their other possible type when choosing a mediator.

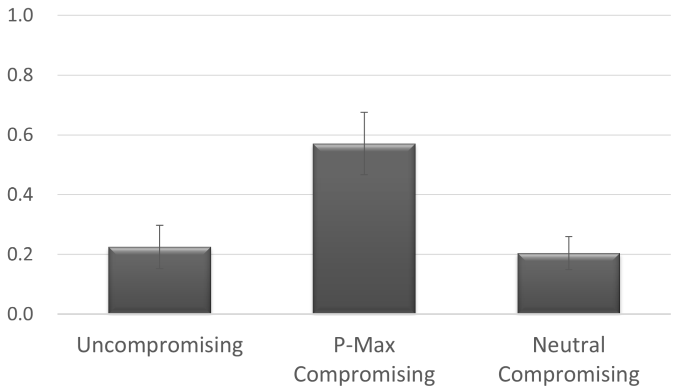


Figure 7: Proportion of Mediator-Selection Rule Chosen (All Principals)

Finding 5. *Consistent with Hypothesis 2(b), the majority of the informed principals choose a Compromising rule in which both types choose the same mediator.*

This finding suggests that the subjects understand that they should not make type-revealing mediator choices (nonequilibrium separating play of the game), in line with the behavioral prediction in our setting. This implies that the empirically-observed type-dependent mediator choices in Figure 3 in Experiment I have little to do with the failure of inscrutable mediator selection. Thus, we interpret Finding 5 as strong evidence for the subjects recognizing the need for inscrutable mediator selection and making some kind of intertype compromise between the goals of their true type and of their other possible type so as to choose inscrutably.[35]

[35] It could be possible that the subjects rarely choose the Uncompromising option not because they anticipate such a choice could signal their private information but because they just do not realize the strategic usefulness of it. We rule out this explanation because we let subjects experience each of the options and they appear to understand the implications of each choice.

Importantly, we observe in Figure 7 that informed principals choose the P-Max Compromising rule significantly more frequently (57.1%) than the Neutral Compromising rule (20.4%).[36] This implies that the majority of informed subjects perceive the P-Max Mediator (that the L type prefers) as a reasonable intertype compromise, counter to the theory of neutral optimum that predicts the Neutral Mediator to be the most reasonable compromise.

To see why the inscrutable intertype compromise gets resolved in favor of the L type, we observe subjects' choices of mediator-selection rules by type. Figure 8 shows the proportions of the H and L types of all principals that choose each of the three mediator-selection rules. Comparing the two bars for P-Max Compromising in Figure 8 to the two bars on the right in panels (a) and (b) of Figure 3, we observe consistent patterns of divergent principal behavior by type. That is, the choice among the two Compromising rules is dependent on the types. In particular, the L types significantly more frequently choose P-Max Compromising (63.0%) over Neutral Compromising (12.9%), whereas the H types more frequently choose Neutral Compromising (45.0%) over P-Max Compromising (37.9%) although the difference is statistically insignificant.

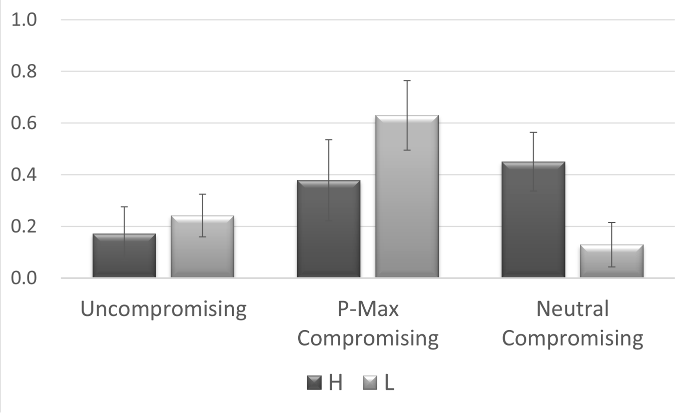
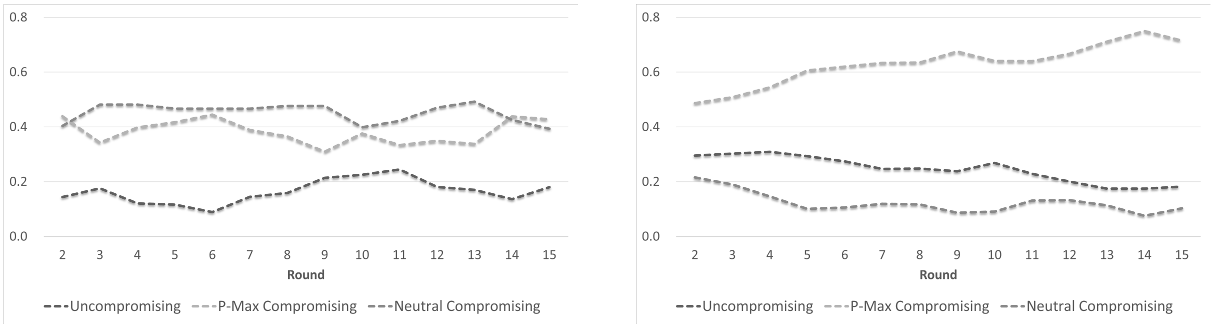


Figure 8: Proportion of Mediator-Selection Rule Chosen by Type (All Principals)

Figure 9 compares principal behavior over time between the two types. The figure presents the 3-round moving averages of the frequencies of three mediator-selection rules chosen conditional on H type and L type.[37] The H types consistently choose Neutral Compromising with the frequencies ranging from 39.3% to 49.2% but there is also a persistent

[36] The Wilcoxon signed-rank test confirms this observation (one-sided p-value< 0.0005).

[37] The moving average for round n is calculated by averaging the frequencies in rounds $n-1$, n, and $n+1$. The data points accordingly start at Round 2 and end at Round 15.



(a) H type

(b) L type

Figure 9: Trends of Frequencies of Mediator-Selection Rules Chosen Conditional on Types (3-Round Moving Averages)

non-trivial fraction of H types choosing P-Max Compromising (31%–44.4%). For L types, the overall tendency is toward P-Max Compromising right from the outset. The frequency of P-Max Compromising chosen by L types gradually rises, starting at around 49% and eventually reaching about 72%.

Either the P-Max or the Neutral Mediator is a reasonable inscrutable intertype compromise for both types of the principal. Excluding those who did not make the inscrutable choice (i.e., chose Uncompromising), the majority of the L types sought the compromise that favored their own interests, whereas roughly half of the H types sought the compromise that favored their own interests while the other half compromised with their own interests. In the aggregate, principals pool on the P-Max Compromising rule, the bulk of which can be attributed to L type behavior. This rationalizes why the concept of neutral optimum fails in the lab. A significant fraction of principals understand correctly that they should be inscrutable in their mediator choice by choosing a Compromising rule that reflects their intertype compromise; but *the two types disagree on how they should make their inscrutable intertype compromise.*[38] As we explained earlier, the concept of neutral optimum predicts that the compromise gets resolved in favor of the H type by L types compromising by choosing the Neutral Mediator. While this behavior must happen instantly through the internal process of compromising at the outset in theory, subjects might at least learn to do so over

[38] We obtain a qualitatively consistent result from the individual data analysis provided in Appendix C.3. At the individual level, a substantial proportion of those who appear to make the intertype compromise fail to reach an agreement about their intertype compromise.

rounds. However, H types consistently do not choose Neutral Compromising as much and L types rather increasingly pool on the P-Max Mediator over rounds.

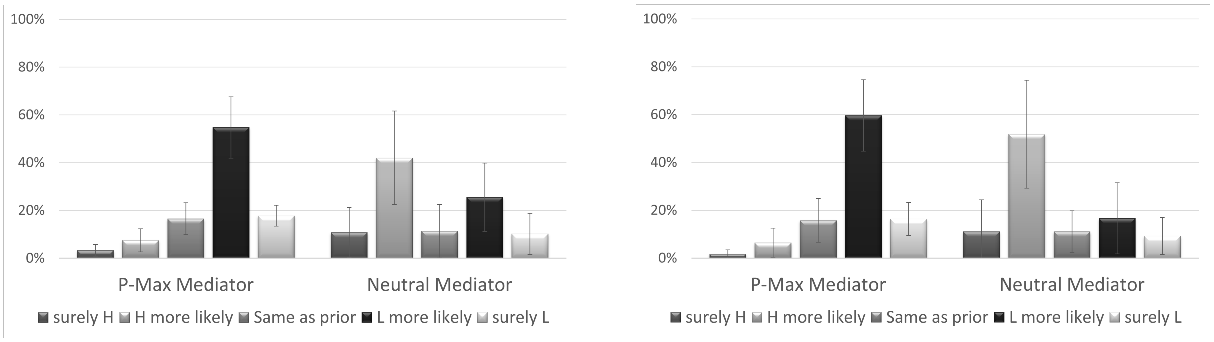
The observed phenomenon in which each type favors a compromising rule that benefits its own type is well explained by *projection bias* (Loewenstein, O'Donoghue and Rabin, 2003). When the H type principal considers the L type's maximization problem to make an intertype compromise, projection bias leads the H type to mispredict the interim utility of the L type as a convex combination of the true interim utility of the L type and the interim utility of the H type. Consequently, the H type concludes that the compromise should favor the H type. Conversely, when the realized type is L, the same projection bias causes the principal to draw an opposite conclusion. The k-means clustering analysis presented in Appendix C.3 demonstrates that the behavior of individuals categorized in Clusters 2 and 3 aligns with this projection bias. Myerson's (1983) set of axioms for the neutral optimum aims to capture the virtual bargaining process involving intertype compromise. However, our data suggest that the actual intertype compromise is influenced by behavioral projection bias, resulting in outcomes that are inconsistent with the neutral optimum.

5.3.2 Subordinate's Inference and Strategy

To better understand subject behavior, we provide additional observations from our data on subjects' inference and strategy in mediation.

After the principal's mediator choice (according to her chosen mediator-selection rule) is announced, a subordinate chooses one of the five possible inferences about the principal's type: (i) Surely H, (ii) More likely to be H, (iii) Same as the prior, (iv) More likely to be L, and (v) Surely L. Figure 10 reports the frequencies of five inferences by subordinates conditional on either P-Max or Neutral Mediator chosen by the principal for all rounds in panel (a) and for the last 5 rounds in panel (b).

Information is again revealed by the mediator choice, and the pattern of inferences is similar to that of Figure 4 in Experiment I. In the data aggregated across all rounds, subordinates infer that the principal is more likely to be L type after observing the principal's choice of P-Max Mediator with a significantly higher frequency (54.8%) than other inferences; Subordinates infer that the principal is more likely to be H type after observing the



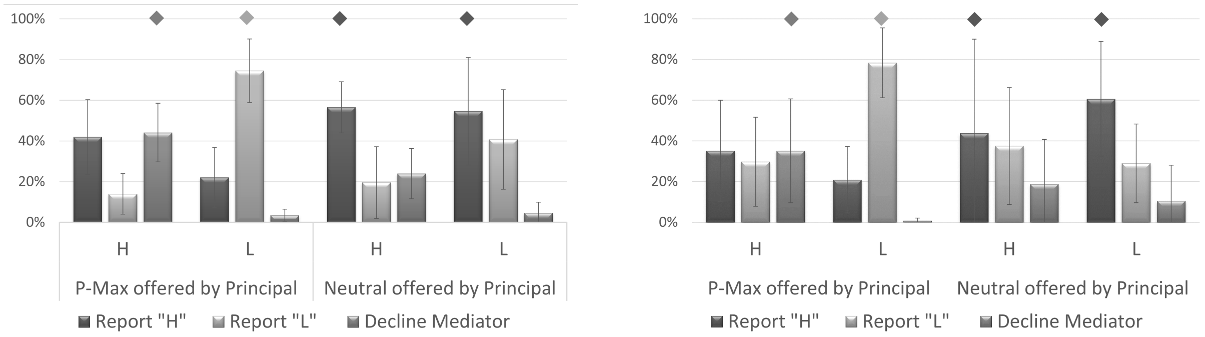
(a) All Rounds

(b) Last 5 Rounds

Figure 10: Subordinate's Inference Conditional on Each Mediator Chosen

principal's choice of Neutral Mediator with a significantly higher frequency (42.0%) than other inferences.[39] A non-trivial fraction of subjects infer that the principal is more likely to be L type given the Neutral mediator chosen (25.6%), but the frequency of "H more likely" inference increases to 51.9% and that of "L more likely" inference decreases down to under 17% in the last 5 rounds, as can be seen in panel (b) of Figure 10.

Figure 11 shows the frequencies of different strategies that subordinates choose by type conditional on either P-Max or Neutral Mediator offered by the principal (according to her chosen mediator-selection rule), aggregated over all rounds in panel (a) and aggregated over the last 5 rounds in panel (b).



(a) All Rounds

(b) Last 5 Rounds

Note: The diamond shapes indicate the subject's best response of each type to the predominant reported belief about the principal's type (that L more likely than the prior given P-Max and H more likely than the prior given Neutral). See Appendix C.2 for the computation of best responses.

Figure 11: Subordinate's Strategy in Chosen Mediator by Type

[39] The Wilcoxon signed-rank tests show one-sided p-value<0.05 for all relevant pairwise comparisons.

The figure in panel (a) above makes clear a number of regularities that are consistent with panel (a) of Figure 5 in Experiment I. One similarity of our particular interest is the rejection behavior.[40] The frequency that H type subordinates decline the P-Max Mediator chosen by the principal is only 44.1%, which even reduces to 35.1% in the last 5 rounds. The H type subordinates also decline the Neutral Mediator chosen by the principal at a low but non-trivial rate of 23.9% in all rounds and of 18.8% in the last 5 rounds. On the other hand, L types hardly decline either P-Max or Neutral.

When subjects are *given* either of the two Compromising rules (Parts R2 and R3), they tend to correctly understand that their mediator choice does not reveal information about their type; whereas when given the Uncompromising rule (Part R1), subjects correctly understand that their mediator choice is type-revealing (see Appendix D.2). When subjects are *choosing* a mediator-selection rule (Part I-RS), they are instructed and reminded that their mediator choice is automatically assigned based on their chosen rule. Given that the subjects choose the Compromising rules significantly more often than the Uncompromising rule (see Figure 7), in principle, the actual mediator choice by the principal should convey no information, and thus the subordinate should draw no inferences about the type of the principal. However, given the mediator choices, subjects make some inferences and act somewhat optimally given their inferences. The observations of subjects' inference and strategy in Experiment II are consistent with Finding 4(2)-(3) in Experiment I.

The observations that the principal's mediator choice revealed information does not imply that the subjects did not choose inscrutably. Our data already confirmed that the two types made the inscrutable intertype compromise but they sought divergent compromises. A possible explanation for why information was revealed is that the subjects understood and expected that different types disagree on their intertype compromise.

In sum, the theory of neutral optimum predicts that the Neutral Mediator is the most reasonable compromise. In order for different types to agree on such a compromise, the principal should recognize that in making her compromise before choosing, the H type could

[40] Regarding reporting strategies, the frequencies of sincere H messages by H types are 41.9% when offered P-Max and 56.5% when offered Neutral, both of which reduce to 35.1% and 43.8% respectively in the last 5 rounds. L types are more sincere in sending L message when offered P-Max at a significant rate of 74.5%, relative to when offered Neutral (40.8%).

actually benefit greatly from revealing its type while the L type should be concealing, so that the compromise gets resolved in favor of the H type. Our observations indicate that not only the subjects did not make compromises in such a way, but also they were unable to pick up this idea over rounds. What is striking is the lack of fully optimal behavior of H types, reflected in the low frequency of declining the P-Max Mediator. This is the key factor in understanding why the theory of neutral optimum did not work in the lab. Over rounds, H type subordinates decline even less often. In turn, L type principals are not able to experience that their type-revealing choice of P-Max Mediator would be detrimental, thus having no need to make an implicit compromise by choosing the Neutral Mediator.

6 Discussion

We experimentally investigate the informed principal's mechanism selection problem. To our knowledge, our paper is the first to attempt a lab experiment on informed principal problems. We find that, in line with the theory, the uninformed subjects choose the ex ante peace-maximizing mediator. However, contrary to the theory of neutral optimum, the informed subjects do not choose the neutral mediator. We confirm that the informed subjects choose inscrutably, each type making some inscrutable intertype compromise. But the two types disagree on how they should make their compromise. Thus, the intertype compromise resulting in compromising on the neutral mediator does not occur and is not learned over time in the lab. Further, our lab data vividly demonstrate that the subjects make inferences given the mediator choice by the principal. Such inferences, however, do not lead them to "revise" their implicit contemplation of intertype compromise before choosing a mediator, largely due to H type subjects not choosing as much their preferred neutral mediator nor declining as much the peace-maximizing mediator when offered. This is particularly striking because the H type could benefit greatly from revealing its type rather than letting the peace-maximizing mediator be selected.

Myerson (1983) considers the principal's mechanism selection as part of a noncooperative game, which can be executed in the lab. A fundamental issue is that it is hard to build such a game in the lab that is completely free of the possibility of revealing information during a mechanism selection process. The challenge for future work will be to design an experiment in which the mediator choice does not depend in *any way* on one's private information so that the choice itself conveys absolutely no information. Specifically, it would be nice to know whether the neutral mediator is selected over the peace-maximizing one in an environment that systematically embeds the truthful implementation of the chosen mediator and directly tests subjects' consideration of intertype compromise.

We discuss some potential extensions of our experiments. The subjects make inferences and strategically choose their actions accordingly, but they do not incorporate those considerations into their initial mediator choice. One possible explanation is that the subjects do not consider their inferences to be perfect, so if they are provided with some information about what others have chosen based on their types, the subjects might be able to act on their inferences that are confirmed by others. Another possible explanation might be the complexity of the game play under mediation. That is, the informed principal's mechanism selection problem already embeds the theory of signaling in markets with adverse selection (or information leakage problem) in addition to the complexity of the mechanism design problem. A crucial question is then how the complexity of the mechanism itself affects the signaling through mechanism choice. We might consider a simpler design in which after the principal selects a mediator, each subject just chooses between "in" and "out"; if both choose "in," then without explicitly making inferences nor choosing reporting strategies, their payoffs are realized according to the truthful implementation of the mediator's plan. Alternatively, we can examine the informed principal's mediator selection problem in a sender-receiver environment only with simple obedience constraints.

References

Balkenborg, Dieter and Miltiadis Makris. 2015. "An Undominated Mechanism for a Class of Informed Principal Problems with Common Values." *Journal of Economic Theory* 157:918–958.

Bercovitch, Jacob and Scott Sigmund Gartner. 2009. New Approches, Methods, and Finding in the Study of Mediation. In *International Conflict Mediation: New Approaches and Findings*, ed. Jacob Bercovitch and Scott S. Gartner. New York: Routledge pp. 1–15.

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

Blume, Andreas, Ernest K. Lai and Wooyoung Lim. 2023. "Mediated Talk: An Experiment." *Journal of Economic Theory* 208:105593.

Casella, Alessandra, Evan Friedman and Manuel Perez Archila. 2024. "On the Fragility of Mediation: Theory and Experimental Evidence." Working Paper.

Cella, Michela. 2008. "Informed Principal with Correlation." *Games and Economic Behavior* 64(2):433–456.

Dosis, Anastasios. 2022. "On the Informed Principal Model with Common Values." *The RAND Journal of Economics* 53(4):792–825.

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(4):529–560.

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

Galanter, Marc. 2004. "The Vanishing Trial: An Examination of Trials and Related Matters in Federal and State Courts." *Journal of Empirical Legal Studies* 1(3):459–570.

Goltsman, Maria, Johannes Hörner, Gregory Pavlov and Francesco Squintani. 2009. "Mediation, Arbitration and Negotiation." *Journal of Economic Theory* 144(4):1397–1420.

Holmström, Bengt and Roger B Myerson. 1983. "Efficient and Durable Decision Rules with Incomplete Information." *Econometrica* 51(6):1799–1819.

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

Kim, Jin Yeub. 2017. "Interim Third-Party Selection in Bargaining." *Games and Economic Behavior* 102:645–665.

Koessler, Frédéric and Vasiliki Skreta. 2016. "Informed Seller with Taste Heterogeneity." *Journal of Economic Theory* 165:456–471.

Koessler, Frédéric and Vasiliki Skreta. 2019. "Selling with Evidence." *Theoretical Economics* 14(2):345–371.

Ledyard, John O. and Thomas R. Palfrey. 1994. "Voting and Lottery Drafts as Efficient Public Goods Mechanisms." *Review of Economic Studies* 61:327–355.

Loewenstein, George, Ted O'Donoghue and Matthew Rabin. 2003. "Projection bias in predicting future utility." *the Quarterly Journal of economics* pp. 1209–1248.

Macqueen, J. 1967. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press.*

Maskin, Eric and Jean Tirole. 1990. "The Principal-Agent Relationship with an Informed Principal: The Case of Private Values." *Econometrica* 58(2):379–409.

Maskin, Eric and Jean Tirole. 1992. "The Principal-Agent Relationship with an Informed Principal, II: Common Values." *Econometrica* 60(1):1–42.

Myerson, Roger B. 1982. "Optimal Coordination Mechanisms in Generalized Principal-Agent Problems." *Journal of Mathematical Economics* 10:67–81.

Myerson, Roger B. 1983. "Mechanism Design by an Informed Principal." *Econometrica* 51(6):1767–1797.

Myerson, Roger B. 1984. "Two-Person Bargaining Problems with Incomplete Information." *Econometrica* 52(2):461–488.

Myerson, Roger B. 1991. *Game Theory: Analysis of Conflict*. Cambridge, M.A.: Harvard University Press.

Mylovanov, Tymofiy and Thomas Tröger. 2012. “Informed-principal Problems in Environments with Generalized Private Values.” *Theoretical Economics* 7:465–488.

Mylovanov, Tymofiy and Thomas Tröger. 2014. “Mechanism Design by an Informed Principal: Private Values with Transferable Utility.” *The Review of Economic Studies* 81(4):1668–1707.

Nishimura, Takeshi. 2022. “Informed Principal Problems in Bilateral Trading.” *Journal of Economic Theory* 204:105498.

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

Selten, Reinhard. 1967. “Die strategiemethode zur erforschung des eingeschr nkt rationale verhaltens im rahmen eines oligopolexperiments.” *Beitr ge zur experimentellen Wirtschaftsforschung* p. 136.

Severinov, Sergei. 2008. “An Efficient Solution to the Informed Principal Problems.” *Journal of Economic Theory* 141:114–133.

Skreta, Vasiliki. 2011. “On the Informed Seller Problem: Optimal Information Disclosure.” *Review of Economic Design* 15:1–36.

Stienstra, Donna J. 2011. “ADR in the Federal District Courts: An Initial Report.” Federal Judicial Center. **URL:** `https://www.fjc.gov/content/adr-federal-district-courts-initial-report`.

Wilkenfeld, Jonathan, Kathleen Young, Victor Asal and David Quinn. 2003. “Mediating International Crises: Cross-National and Experimental Perspectives.” *Journal of Conflict Resolution* 47(3):279–301.

A Non-parametric Test Results

Table 5: **Non-parametric Tests for Experiment I (Part 4 in the Simple and Complex Treatments)**

Reference	Test	Null Hypothesis	p-values	
			Simple	Complex
Finding 1	WSR	1.1 (Uninformed, all subjects) The rate of choosing P-Max = the rate of choosing Neutral.	0.125	0.125
	WSR	1.2 (Uninformed, principals) The rate of choosing P-Max = the rate of choosing Neutral.	0.125	0.125
Finding 2	WSR	1.3 (Informed, all subjects) The rate of choosing P-Max = the rate of choosing Neutral.	0.125	0.125
	WSR	1.4 (Informed, principals) The rate of choosing P-Max = the rate of choosing Neutral.	0.125	0.125
Fig. 2	MW	2.1 (all subjects) The rate of choosing P-Max is the same across Uninformed and Informed.	0.0286	0.0571
	MW	2.2 (principals) The rate of choosing P-Max is the same across Uninformed and Informed.	0.0286	0.0286
Fig. 16	MW	2.3 (all subjects/principals) The rate of actually playing P-Max is the same across Uninformed and Informed.	0.0286	0.4596
Fig. 3	WSR	3.1 (Uninformed, all subjects) The rate of choosing P-Max is the same across the two types.	0.125	0.125
Finding 3	WSR	3.2 (Informed, all subjects) The rate of choosing P-Max is the same across the two types.	0.125	0.125
Fig. 3	WSR	3.3 (Informed, all subjects) The rate of choosing P-Max = the rate of choosing Neutral, by H type.	0.25	0.25
	WSR	3.4 (Informed, all subjects) The rate of choosing P-Max = the rate of choosing Neutral, by L type.	0.125	0.125

- ■ WSR and MW refer to the Wilcoxon signed-rank test and Mann-Whitney test, respectively. All null hypotheses are two-sided.
- ■ P-Max and Neutral refer to the P-Max and Neutral Mediators, respectively. Treatment names are abbreviated to Uninformed and Informed.
- ■ In parenthesis, we indicate which treatment data we consider and/or whether we consider data for all subjects or principal-subjects.
- ■ Note that with four independent observations (sessions) in Experiment I, the minimum attainable p-value is 0.125 (two-sided).
- ■ For 4.1–7.4, we test for only subordinate-subjects.

Table 5: **(Continued)**

Reference	Test	Null Hypothesis	p-values	
			Simple	Complex
Fig. 4	WSR	4.1 (Informed) The rate of 'L-inference' given P-Max = the rate of 'L-inference given Neutral.	0.125	0.125
Fig. 18	WSR	4.2 (Uninformed) The rate of 'L-inference' given P-Max = the rate of 'L-inference' given Neutral.	0.125	0.125
	WSR	4.3 (Uninformed) The rate of 'same inference' given P-Max = the rate of 'same inference' given Neutral.	0.125	0.625
Figs.4&18	MW	4.4 The rate of 'same inference' given P-Max is the same across Uninformed and Informed.	0.0286	0.0571
	MW	4.5 The rate of 'same inference' given Neutral is the same across Uninformed and Informed.	0.0286	0.0286
Fig. 4	WSR	5.1 (Informed) Given P-Max chosen, the rate of 'L-inference' = the rate of 'same inference.'	0.125	0.125
	WSR	5.2 (Informed) Given P-Max chosen, the rate of 'L-inference' = the rate of 'H-inference.'	0.125	0.125
	WSR	5.3 (Informed) Given Neutral chosen, the rate of 'H-inference' = the rate of 'same inference.'	0.125	0.125
	WSR	5.4 (Informed) Given Neutral chosen, the rate of 'H-inference' = the rate of 'L-inference.'	0.125	0.125
Fig. 5	WSR	6.1 (Informed) The rate of rejecting P-Max = the rate of rejecting Neutral.	0.625	0.125
	WSR	6.2 (Informed) The rate of rejecting P-Max = the rate of rejecting Neutral, by H type.	0.125	0.125
	WSR	6.3 (Informed) The rate of rejecting P-Max = the rate of rejecting Neutral, by L type.	0.875	0.125
Fig. 6	WSR	6.4 (Uninformed) The rate of rejecting P-Max = the rate of rejecting Neutral.	0.625	0.125
	WSR	6.5 (Uninformed) The rate of rejecting P-Max = the rate of rejecting Neutral, by H type.	0.875	1
	WSR	6.6 (Uninformed) The rate of rejecting P-Max = the rate of rejecting Neutral, by L type.	0.25	0.625
Figs. 5&6	MW	7.1 The rate of rejecting P-Max by H type is the same across Uninformed and Informed.	0.8857	0.8857
	MW	7.2 The rate of rejecting Neutral by H type is the same across Uninformed and Informed.	0.3429	0.2
	MW	7.3 The rate of rejecting P-Max by L type is the same across Uninformed and Informed.	0.2	0.8857
	MW	7.4 The rate of rejecting Neutral by L type is the same across Uninformed and Informed.	0.6857	0.34286

■ WSR and MW refer to the Wilcoxon signed-rank test and Mann-Whitney test, respectively. All null hypotheses are two-sided.
■ P-Max and Neutral refer to the P-Max and Neutral Mediators, respectively. Treatment names are abbreviated to Uninformed and Informed.
■ In parenthesis, we indicate which treatment data we consider and/or whether we consider data for all subjects or principal-subjects.
■ Note that with four independent observations (sessions) in Experiment I, the minimum attainable p-value is 0.125 (two-sided).
■ For 4.1–7.4, we test for only subordinate-subjects.

Table 6: **Non-parametric Tests for Experiment II (Part 4)**

Reference	Test	Null Hypothesis	p-values
Fig. 7 Finding 5	WSR	1.1 The rate of choosing P-Max Comp. = the rate of choosing Uncompromising.	<0.001
	WSR	1.2 The rate of choosing P-Max Comp. = the rate of choosing Neutral Comp.	<0.001
	WSR	1.3 The rate of choosing Compromising (both P-Max and Neutral) = the rate of Uncompromising.	0.0313
Fig. 8	WSR	2.1 For H type, the rate of choosing P-Max Comp. = the rate of choosing Uncompromising.	0.037
	WSR	2.2 For H type, the rate of choosing Neutral Comp. = the rate of choosing Uncompromising.	0.0022
	WSR	2.3 For H type, the rate of choosing P-Max Comp. = the rate of choosing Neutral Comp..	0.8923
	WSR	2.4 For L type, the rate of choosing P-Max Comp. = the rate of choosing Uncompromising.	<0.001
	WSR	2.5 For L type, the rate of choosing Neutral Comp. = the rate of choosing Uncompromising.	0.98992
	WSR	2.6 For L type, the rate of choosing P-Max Comp. = the rate of choosing Neutral Comp.	<0.001
Fig. 10(a)	WSR	3.1 Under P-Max Comp., the rate of 'L more likely' inference = the rate of 'Surely H' inference.	0.003
	WSR	3.2 Under P-Max Comp., the rate of 'L more likely' inference = the rate of 'H more likely' inference.	<0.001
	WSR	3.3 Under P-Max Comp.,the rate of 'L more likely' inference = the rate of 'Same as prior' inference.	<0.001
	WSR	3.4 Under P-Max Comp., the rate of 'L more likely' inference = the rate of 'Surely L' inference.	<0.001
	WSR	3.5 Under Neutral Comp., the rate of 'H more likely' inference = the rate of 'Surely H' inference.	0.0082
	WSR	3.6 Under Neutral Comp., the rate of 'H more likely' inference = the rate of 'Same as prior' inference.	0.0082
	WSR	3.7 Under Neutral Comp., the rate of 'H more likely' inference = the rate of 'L more likely' inference.	0.065
	WSR	3.8 Under Neutral Comp., the rate of 'H more likely' inference = the rate of 'Surely L' inference.	0.0028
Fig. 11(a)	WSR	4.1 The rate of rejecting P-Max Mediator = the rate of rejecting Neutral Mediator	0.2188
	WSR	4.2 The rate of rejecting P-Max Mediator = the rate of rejecting Neutral Mediator, by H type.	0.0313
	WSR	4.3 The rate of rejecting P-Max Mediator= the rate of rejecting Neutral Mediator, by L type.	0.3613

■ WSR refers to the Wilcoxon signed-rank test. All null hypotheses are two-sided.

■ Comp. refers to Compromising. For 3.1–3.8, P-Max Mediator is announced under P-Max Comp., and Neutral Mediator is announced under Neutral Comp.

■ For 4.1–4.3, we test the rates of rejecting P-Max or Neutral Mediator offered by the principal according to her chosen mediator-selection rule.

Online Appendix for

"Toward an Understanding of Optimal Mediation Choice"

Jin Yeub Kim Wooyoung Lim

Appendix B provides discussions of our modeling choices, and the relevant theoretical characterizations and results. Appendix C gives additional analyses for the experimental results. Appendix D contains additional figures.

B Supplemental Materials for Section 2

B.1 Justifications for the Baseline Environment

Our baseline environment can be formulated as a Baysian bargaining problem à la Myerson (1984), where a set of possible outcomes is given instead of a set of actions or strategies for each player. With the restriction that the only possible peaceful settlement is an equal split, the set of possible outcomes available to the two players is $D = \{d^a, d^*\}$. Here, d^a represents the agreement outcome (the equal split) that is jointly feasible for the players together and d^* represents the disagreement outcome (war) that will occur if the players fail to cooperate or fail to agree on a mediator. For each player i, $T_i = \{H, L\}$ is the set of possible types t_i for player i. As in the main text, q is the probability of the H type, which is common for both players. Let $T = T_1 \times T_2$ denote the set of all possible type combinations $t = (t_1, t_2)$. The payoffs that each player would get if $d \in D$ were chosen and if t were the vector of the players' types can all be specified.

In this problem, a mediator (represented by its mechanism) specifies how the choice $d \in D$ should depend on the types $t \in T$ reported by the players. We restricted attention to symmetric mechanisms, and denoted by p_H, p_M, and p_L for the mediator's probabilities of recommending the agreement outcome if the reported types were (h, h), (h, l) or (l, h), and (l, l), respectively.

There are two incentive constraints relevant to this setup.[41] First, a player might be tempted to lie about her type to the mediator. The mechanism (p_H, p_M, p_L) is (Bayesian)

[41] The incentive constraints in Bayesian bargaining problems are extensively discussed in Myerson (1984, p.464) and Myerson (1991, pp.263-267).

incentive compatible iff it satisfies the following informational incentive constraints for the H and L types, respectively:

$$\begin{aligned}
&q(p_H(1/2)+(1-p_H)\theta/2)+(1-q)(p_M(1/2)+(1-p_M)\delta\theta)\\
&\geq q(p_M(1/2)+(1-p_M)\theta/2)+(1-q)(p_L(1/2)+(1-p_L)\delta\theta);\\
&q(p_M(1/2)+(1-p_M)(1-\delta)\theta)+(1-q)(p_L(1/2)+(1-p_L)\theta/2)\\
&\geq q(p_H(1/2)+(1-p_H)(1-\delta)\theta)+(1-q)(p_M(1/2)+(1-p_M)\theta/2).
\end{aligned} \tag{1}$$

The revelation principle applies: Myerson (1982) shows that there is no loss of generality in considering only incentive compatible mechanisms that satisfy (1).[42] Second, the players get the disagreement outcome if they fail to cooperate, and any player can force it whenever it might be profitable. That is, in our situation, a mechanism cannot be implemented without the prior agreement of the players. The mechanism (p_H, p_M, p_L) is *individually rational* iff it satisfies the following participation constraints for the H and L types, respectively:

$$\begin{aligned}
&q(p_H(1/2)+(1-p_H)\theta/2)+(1-q)(p_M(1/2)+(1-p_M)\delta\theta)\\
&\geq q\theta/2+(1-q)\delta\theta;\\
&q(p_M(1/2)+(1-p_M)(1-\delta)\theta)+(1-q)(p_L(1/2)+(1-p_L)\theta/2)\\
&\geq q(1-\delta)\theta+(1-q)\theta/2.
\end{aligned} \tag{2}$$

We can, with no loss of generality, assume that the players will agree on a mechanism that satisfies (2) for all types.[43] We say the mechanism (p_H, p_M, p_L) is *(incentive) feasible* if it satisfies (1) and (2). We assume that the players only consider choosing among feasible mechanisms.

We can equally formulate this problem as a general Bayesian incentive problem á la Myerson (1983, pp.1769-1772). Let $D_0 = \{d^a, d^*\}$ be the set of all possible *enforceable actions*. For each player i, let D_i be the set of all possible *private actions* that are privately

[42] See the relevant discussions in Myerson (1983, p.1772; 1984, p.464; 1991, p.264).

[43] Myerson (1991) asserts that "[g]iven any equilibrium of a mechanism-selection game in which some players would sometimes insist on the disagreement outcome, there is an equivalent individually rational mechanism that would choose the disagreement outcome whenever one or more players would insist on it in the given equilibrium of the mechanism-selection game" (p.267).

controlled by player i. We redefine $D = D_0 \times D_1 \times D_2$, with d denoting a typical outcome in D. In this problem, a mechanism would choose an outcome $d = (d_0, d_1, d_2) \in D$ as a function of types reported. Then the enforceable action d_0 is carried out, and each player i is confidentially informed that d_i is the private action recommended for her. We can define the incentive constraints that give the players incentives to report their types honestly and carry out their recommended private actions obediently when the mechanism is implemented. Our baseline environment is a special case of this problem, where each player's set of private actions is simply $D_i = \{$"accept","reject"$\}$ and all players get the war payoffs if any player chooses her "reject" option. The d^* is an enforceable action that also gives the war payoffs. Then we can restrict attention without loss to mechanisms in which no player is ever asked to "reject" because the d^* action may be used instead. So the incentive constraints reduce to (1) and (2), which ensure that no player has any incentive to lie or reject in the mediation.

In the Nash demand game, the players bargain over how to split the surplus. HMS take this bargaining process to the mediation game so that, given type reports, the mediator publicly recommends a split $(x, 1-x)$ or war. Each player can then separately decide whether to accept or reject the recommended split. Unless both players accept, war takes place. Taking this mediation protocol to the lab, CFP constrain the mediator's recommendations to lie in a restricted set containing only those that appear in the optimal mediation mechanism: $\{(1-\theta, \theta), (1/2, 1/2), (\theta, 1-\theta), w\}$ where w stands for "walking out," which is equivalent to "war recommendation." We constrain them further to $\{(1/2, 1/2), w\}$. This restriction serves two important purposes in our paper.

First, with this restriction, our mediation protocol can simply be stipulated without a stage where the players decide whether to accept the mediator's recommendations, as if it were an arbitration protocol. But this does not necessarily mean that our mediator has enforcement power (which HMS call the arbitrator). Why? When the mediator's recommendation of peaceful settlement is restricted to an equal split, there is no difference in optimal recommendation strategies between the mediator with enforcement power and the mediator without enforcement power.[44] When defining incentive feasibility for mediators

[44] In HMS, the mediator (without enforcement power) can effectively circumvent the unenforceability constraint by using recommendation strategies that do not reveal to a disputant that the opponent is weak, i.e., obfuscation.

without enforcement power (i.e., who can only make non-binding recommendations), we need two (new) incentive compatibility constraints with double deviation and two ex post participation constraints, as defined by HMS. But with only one agreement outcome possible, Proposition 1 of Kim (2017) proves that the arbitration and mediation feasible mechanisms sets are equivalent. Accordingly, the arbitrator's optimal recommendation of the equal split would be self-enforcing.[45] Thus, unlike HMS, the discussion of whether our mediator has enforcement power or not is immaterial in our context, and our purpose is not comparing mediation to arbitration. Note, however, that the subordinate in the mediator selection game has an option to go to war (or decline the mediator, as phrased in the lab) after the principal's chosen mediator is announced. This is about an outcome that any player can force when the announced mediator is to be played, which is taken care of by the participation constraints (2), and is not about disobeying or rejecting the mediator's recommendations.

Second, the restriction allows us to simplify the representation of mediators. If a mediator can recommend different splits $(x, 1-x)$, $x \in [0, 1]$, then the description of mediator will involve the split recommendation under agreement as well as the probability of recommending agreement, as functions of type reports. Instead, given that the only possible split recommendation is $(1/2, 1/2)$, each mediator can be represented only by the probability of recommending agreement given type reports.

B.2 Justifications for the Mediator Selection Game

In our conflict environment, war is an outcome that will occur by default if the players fail to cooperate or fail to agree on a mediator. In Appendix B.1, we have formulated our environment as a Bayesian bargaining problem where war is the disagreement outcome. In such a problem, a feasible mediator must satisfy the participation constraints (2) that ensure every player would agree to participate in the mediation rather than forcing the disagreement out-

[45] To briefly explain the intuition, our participation constraints (2) ensure that the incentive feasible mechanism probabilities (p_H, p_M, p_L) are chosen so that these probabilities induce posterior beliefs of the players that make them willing to accept the equal split if recommended, beyond making them better off in expectation by participating in mediation; and together with the incentive compatibility constraints (1), the players subsequently have no choice but to honestly report their types and to "voluntarily" obey the equal-split recommendation of the mediator. In particular, the mediator's recommendation of the equal split does not reveal that one player is L type to an H type opponent; so the mediator who "prescribes" the equal split is essentially not using its enforcement power but rather such a prescription is self-enforcing.

come. Because the principal chooses among feasible mediators, her participation constraints are already satisfied at stage 2 of the mediator selection game given her prior beliefs about the subordinate's type. So when the principal's announced mediator is to be played, the principal only decides whether to send a truthful message to the mediator given no new information. But because the subordinate may make some inferences about the principal's type based on the announcement, his participation constraints for the chosen mediator might be violated given his updated beliefs about the principal's type. Thus, in implementing the announced mediator, the subordinate must decide concurrently whether to go to war or to participate, and if he chooses to do the latter, what message to send, given his (possibly new) information. Here, the subordinate's "participate" would be implied by sending a message to the mediator. So we set up our mediator selection game so that only the subordinate has an additional option of going to war instead of reporting its type in stage 2.

One might consider separating out the subordinate's decisions in stage 2 as follows: In the second stage, the subordinate chooses either to go to war (rejecting the mediator) or to participate in the mediation. If he agrees to participate, then the two players play the mediator in the third stage, with each player confidentially reporting its type to the mediator. This game resembles Maskin and Tirole's (1992) three-stage mechanism selection game. In their setting, the subordinate does not have private information, so only the subordinate would update his beliefs about the principal's type based on the principal's choice in the first stage. However, the subordinate has private information in our setting, so the principal may be able to infer something about the subordinate's type from his participation decision in the second stage, in addition to the possibility that the subordinate may infer something about the principal's type from her choice in the first stage. Thus, using the three-stage game would only add another layer of the information leakage problem, complicating the informed principal's mediator selection problem both in theory and in the lab without adding commensurate insights. Our two-stage game described in Section 2.2 allows us to focus on the informed principal's dilemma regarding her private information.

B.3 Mediator Characterization

To describe our theoretical characterizations, we use the following three statistics, the first two of which were used in HMS: $\lambda \equiv \frac{q}{1-q}$, $\gamma_H \equiv \frac{\delta\theta - 1/2}{1/2 - \theta/2}$, and $\gamma_L \equiv \frac{1/2-\theta/2}{1/2-(1-\delta)\theta}$. The parameter $\lambda > 0$ is the H/L type odds ratio; $\gamma_H > 0$ measures the H type's net benefit of war against an L type relative to its net cost of war against an H type, and $\gamma_L > 0$ measures the L type's net benefit of agreement with an L type relative to that with an H type. With this simplification, the assumption of $q\theta/2 + (1-q)\delta\theta > 1/2$ can be rewritten as $\lambda < \gamma_H$.

Efficient Mechanisms and Peace-Maximizing Mechanism We apply the concept of efficiency to further identify the set of mediators that the players could reasonably consider to choose from among the incentive-feasible ones. The proper concept of efficiency is *interim incentive efficiency* for games in which the players already know their private information when the game begins; and is *ex ante incentive efficiency* for games in which the players learn their private information during the game (Holmström and Myerson, 1983).

We first characterize the set of all interim incentive efficient mechanisms in the following proposition, adapted from Proposition 2 in Kim (2017) stated here without proof.

Proposition 1 (Kim 2017)**.** *For the baseline model with $\gamma_H > \gamma_L$, any mediation mechanism (p_H, p_M, p_L) that satisfies the following characteristics is interim incentive efficient (IIE):*

1. *For $\lambda < \gamma_L$, $p_L = 1$, $p_M \in [0, \lambda/\gamma_H]$, $p_H = 1$.*
2. *For $\gamma_L \le \lambda < \gamma_H$, $p_L = 1$, $p_M \in \left[0, \frac{\gamma_L}{\gamma_L + \gamma_H - \lambda}\right]$, $p_H = p_M + (1-p_M)\gamma_L/\lambda$.*

Focusing on the model with $\gamma_H > \gamma_L$ does not lose the model's generality. In fact, the characterization of IIE mechanisms for the model with $\gamma_H \le \gamma_L$ is subsumed by Case 1 in Proposition 1.[46] Note that p_L is always one, and p_H is determined given p_M and is increasing in p_M. Therefore, we can use p_M as the sole parameter that represents each IIE mediator.

Given the experimental parameter values $\theta = 0.75$, $\delta = 0.8$, and $q = 1/4$ or $2/5$, the IIE mediators can be characterized by Proposition 1 as follows:

1. For $q = 1/4$, $p_L = 1$, $p_M \in [0, 5/12]$, and $p_H = 1$.

[46] For when $\gamma_H \le \gamma_L$, the upper bound on p_M is λ/γ_H if $\lambda < \gamma_H$ and 1 if $\gamma_H \le \lambda < \gamma_L$.

2. For $q = 2/5$, $p_L = 1$, $p_M \in [0, 75/103]$, and $p_H = 15/28 + (13/28)p_M$.

The following proposition characterizes the ex ante incentive efficient mechanism.

Proposition 2. *For the baseline model with $\gamma_H > \gamma_L$, there is a unique ex ante incentive efficient mechanism such that*

1. *For $\lambda < \gamma_L$, $p_L = 1$, $p_M = \lambda/\gamma_H$, $p_H = 1$.*
2. *For $\gamma_L \le \lambda < \gamma_H$, $p_L = 1$, $p_M = \frac{\gamma_L}{\gamma_L + \gamma_H - \lambda}$, $p_H = p_M + (1 - p_M)\gamma_L/\lambda$.*

Proof. Because ex ante incentive efficiency implies interim incentive efficiency (Holmström and Myerson, 1983), we solve for the ex ante incentive efficient mechanism among all IIE mechanisms. A player's ex ante expected utility in mechanism $p \equiv (p_H, p_M, p_L)$, denoted by $U(p)$, is:

$$\begin{aligned} U(p) &\equiv q^2(p_H(1/2) + (1 - p_H)(\theta/2)) + q(1 - q)(p_M(1/2) + (1 - p_M)(\delta\theta)) \\ &\quad + (1 - q)q(p_M(1/2) + (1 - p_M)(1 - \delta)\theta) + (1 - q)^2(p_L(1/2) + (1 - p_L)(\theta/2)) \\ &= (1/2 - \theta/2)Q(p) + \theta/2, \end{aligned}$$

where $Q(p) \equiv q^2 p_H + 2q(1-q)p_M + (1-q)^2 p_L$, which is the ex ante probability of agreement in mechanism $p \equiv (p_H, p_M, p_L)$ given q. Hence, the optimization problem of maximizing $U(p)$ differs from that of maximizing $Q(p)$ only by a positive linear transformation. Note that for any given IIE mechanism, $p_L = 1$ and p_H is increasing in p_M. So we can easily see that the mechanism that maximizes $Q(p)$ is the one that has the highest value of p_M among all IIE mechanisms. Thus it is the unique ex ante incentive efficient mechanism. □

The ex ante incentive efficient mechanism is the IIE mediator with the highest possible p_M, thus associated with the highest ex ante probability of agreement among all IIE mediators. We labeled such a mediator as "P-Max Mediator," which corresponds to the optimal mediation program studied in HMS and used in CFP.[47]

[47] For the class of models considered in HMS, CFP, and this paper, achieving ex ante incentive efficiency is equivalent to maximizing the ex ante probability of peace.

Neutral Mechanism The set of IIE mediators is quite large; Corollary 1 in Kim (2017) implies that the P-Max Mediator is the best feasible mechanism for an L type player, whereas the IIE mediator with the lowest possible p_M (which we labeled as the Neutral Mediator) is the best feasible mechanism for an H type player, among all IIE mediators. When the feasible mechanism that is best for each player depends on what her private type is in such a way, the player cannot choose (and implement) the one that is best for her unless the other player believes that both types would have inscrutably selected the same mechanism without sharing any information during the selection process. Otherwise, the selection of the mechanism itself may convey information about her type to her opponent, and with this new information, the opponent may find new opportunities to gain by dishonesty or forcing the disagreement outcome (see Appendix B.5). Then to be inscrutable, the player must choose the one that will be perceived as a reasonable compromise between the different goals of her different possible types, so as to prevent the other player from learning her type.

Myerson (1983) develops several notions of what such a reasonable *inscrutable intertype compromise* should be. Among those notions, the concept of neutral optimum is a powerful solution concept that identifies the player's inscrutable mechanism uniquely in our setting. The neutral optimum is defined as an incentive-feasible mechanism that cannot be blocked with any concept of blocking that satisfies four axioms, which we do not scrutinize here. The key properties of the neutral optimum are that the solution must be an inscrutable intertype compromise and that it eliminates some mechanisms that would be unreasonable selections for some types during a mechanism-selection process because some players would choose to reveal information about their types rather than let these mechanisms be selected. That is, the concept of neutral optimum refines how the informed principal should make the inscrutable intertype compromise in the stronger sense than does (Bayesian) sequential rationality in the concept of sequential equilibrium.

For this paper's setting, Proposition 3 in Kim (2017) establishes the characterization of neutral mechanism, stated here without proof.

Proposition 3 (Kim 2017)**.** *For the baseline model with $\gamma_H > \gamma_L$, there is a unique neutral mechanism such that*

1. *For $\lambda < \gamma_L$, $p_L = 1$, $p_M = 0$, $p_H = 1$.*

2. *For $\gamma_L \leq \lambda < \gamma_H$, $p_L = 1$, $p_M = 0$, $p_H = \gamma_L/\lambda$.*

The neutral mechanism is the IIE mediator with the lowest possible p_M among all IIE mediators, which we labeled as "Neutral Mediator."

B.4 Alternative Mediator Choice

We have imposed the concept of Pareto efficiency to characterize possible options of mediators that the players can consider to bring to the mediation table. Maskin and Tirole (1992) (henceforth, MT92) also consider the problem of mechanism selection by an informed principal taking a different approach from Myerson (1983). In their analysis, the *Rothschild-Stiglitz-Wilson (RSW) allocation* plays a crucial role.[48] MT92 characterize the equilibrium set of the mechanism selection game, which consists of the allocations that weakly Pareto dominate the RSW allocation. So the RSW allocation can be thought of as the worst equilibrium for every type of principal. MT92 prove that the RSW allocation is the unique allocation that passes the intuitive criterion under the assumption that only the principal has private information.[49] However, as noted by MT92, the RSW allocation may not be interim efficient relative to the prior beliefs (or to any strictly positive beliefs) and there are many equilibria allocations that are not even weakly interim efficient as well.

We characterize the RSW allocation (call it the "RSW Mediator") for our setting.

Proposition 4. *For the baseline model with $\gamma_H > \gamma_L$, the RSW Mediator is characterized by $p_L = 1$, $p_M = p_H = 0$.*

Proof. The RSW allocation is defined to be an allocation that each type of the principal maximizes her own utility within the set of allocations that are incentive compatible (for the principal) and, regardless of the principal's type, yields the subordinate at least his reservation utility. In our setting, the constraints translate to the principal's (interim) feasibility constraints and the subordinate's *ex post* feasibility constraints which assert that the subordinate is willing to play the mechanism and report truthfully when he knows the principal's

[48] The formal definition of the RSW allocation can be found in MT92. The RSW allocation is the best *safe mechanism* (Myerson, 1983).

[49] Nishimura (2022) extends this result to the trading environment with bilateral asymmetric information.

true type. Because the incentive structure is the same for both principal and subordinate, the subordinate's ex post feasibility constraints imply the principal's interim feasibility constraints. Therefore, the RSW Mediator is the solution to the following program:

$$\text{For all } t \in \{H, L\}, \quad \max_{p=(p_H,p_M,p_L)} U_1(p|t)$$

$$\begin{aligned}
\text{subject to: } & p_H(1/2) + (1-p_H)\theta/2 \geq p_M(1/2) + (1-p_M)\theta/2, && \text{(HH-EPIC)}\\
& p_M(1/2) + (1-p_M)(1-\delta)\theta \geq p_H(1/2) + (1-p_H)(1-\delta)\theta, && \text{(LH-EPIC)}\\
& p_M(1/2) + (1-p_M)\delta\theta \geq p_L(1/2) + (1-p_L)\delta\theta, && \text{(HL-EPIC)}\\
& p_L(1/2) + (1-p_L)\theta/2 \geq p_M(1/2) + (1-p_M)\theta/2, && \text{(LL-EPIC)}\\
& p_H(1/2) + (1-p_H)\theta/2 \geq \theta/2, && \text{(HH-EPIR)}\\
& p_M(1/2) + (1-p_M)(1-\delta)\theta \geq (1-\delta)\theta, && \text{(LH-EPIR)}\\
& p_M(1/2) + (1-p_M)\delta\theta \geq \delta\theta, && \text{(HL-EPIR)}\\
& p_L(1/2) + (1-p_L)\theta/2 \geq \theta/2, && \text{(LL-EPIR)}
\end{aligned}$$

where $U_1(p|t)$ is the t-type principal's expected utility in the mediator with $p = (p_H, p_M, p_L)$. That is, $U_1(p|H) = q(p_H(1/2)+(1-p_H)\theta/2)+(1-q)(p_M(1/2)+(1-p_M)\delta\theta)$ and $U_1(p|L) = q(p_M(1/2) + (1-p_M)(1-\delta)\theta) + (1-q)(p_L(1/2) + (1-p_L)\theta/2)$. The HH-EPIC and LH-EPIC constraints together imply $p_H = p_M$. The HL-EPIC and LL-EPIC constraints imply $p_L \geq p_M$. The four EPIR constraints imply, respectively, $p_H \geq 0$, $p_M \geq 0$, $p_M = 0$, and $p_L \geq 0$. Thus, the RSW Mediator must have $p_H = p_M = 0$. Now $U_1(p|L) = q(1-\delta)\theta + (1-q)(p_L(1/2-\theta/2)+\theta/2)$ is maximized by setting $p_L = 1$. □

The RSW Mediator prescribes peace only when the reported types are both L. The associated ex ante probability of peace is $1/4 = 0.25$. The RSW Mediator is not only interim incentive *inefficient* but also worse than the best separating equilibrium of the unmediated communication game (see Proposition 5 in Appendix C.1). Thus the equilibrium set of mediators that weakly Pareto dominate the RSW Mediator gives too large a set of predictions for our informed mediator selection problem.

Subjects in our experiments have a choice set that is restricted to the two IIE mediators. Any IIE mediator can be supported as a sequential equilibrium of the informed principal's mechanism selection game. Hence, adding a third option of mediator that is also on the IIE frontier (or even letting subjects choose from the whole set of IIE mediators) may only unnecessarily enlarge the choice set for subjects without substantively affecting the results. Alternatively, we may move away from the IIE frontier and add the RSW Mediator as a third option for subjects to choose other than the two IIE mediators. But because the RSW Mediator is worse than the other two for both types of the principal, we conjecture that subjects would hardly ever entertain such an option.

B.5 Separating Equilibrium

By the revelation principle, we may restrict attention without loss of generality to equilibria in which the players participate in the mediation and truthfully report their types to the mediator. Truthfully implementable mediators are identified with two sets of constraints (1) and (2). We show that there is no separating equilibrium in which different types of the principal choose distinct IIE mediators followed by the players' participation and truthful type revelation.

Suppose that there are two IIE mediators μ_H and μ_L, such that the H type principal is expected to choose μ_H and the L type principal is expected to choose μ_L.[50] The principal would choose these mediators in this way only if they satisfy $U_1(\mu_t|t) \geq U_1(\mu_{t'}|t)$ for each $t = H, L$ and $t' \neq t$, and are incentive feasible for the principal (where $U_1(\cdot|t)$ denotes the principal's interim expected payoff in implementing the mediator given that her type is t). By Corollary 1 in Kim (2017), the conflicting incentives of the two different types are well-defined in terms of their opposite preference orderings over all IIE mediators. Thus, we can focus on separating equilibria, if they exist, in which the principal chooses an IIE mediator μ_L such that $p_M \neq 0$ when her type is L and chooses an IIE mediator μ_H such that its p_M is lower than that for μ_L when her type is H.

If the subordinate expects that the principal would choose a mediator in this way, then the chosen mediator could be successfully implemented on the equilibrium path only if it were incentive feasible given the information revealed about the principal's type. That is, because the subordinate would rationally infer that the principal's type is $t = H, L$ when μ_t

[50] Our argument here can be extended to cover randomized mediator-selection.

is chosen, each μ_t must be incentive feasible when the subordinate knows that the principal's type is t.

When μ_H is chosen and announced, then the subordinate would infer that the principal was the H type, so a low-type subordinate would not report truthfully because

$$\begin{aligned} & p_M(1/2) + (1-p_M)(1-\delta)\theta < p_H(1/2) + (1-p_H)(1-\delta)\theta \\ \leftrightarrow\ & (p_M - p_H)(1/2) < (p_M - p_H)(1-\delta)\theta, \end{aligned}$$

where $p_M < p_H$ for any IIE mediator characterized in Proposition 1, and $1/2 > (1-\delta)\theta$, violating the L type's incentive compatibility constraint given the updated belief. Also when μ_L is chosen and announced, then the subordinate would infer that the principal was the L type, so an H type subordinate would have an incentive to refuse to participate in the mediation and choose war instead because

$$p_M(1/2) + (1-p_M)\delta\theta < \delta\theta$$

for $p_M \neq 0$, violating the H type's individual rationality constraint given the updated belief. That is, the chosen mediator, either μ_H or μ_L, becomes infeasible as soon as it is selected. This proves that there is no separating sequential equilibrium of the mediator selection game in which different types choose different IIE mediators followed by the players' participation and truthful reporting in mediation.

C Additional Analyses for Experimental Results

C.1 Experiment I: Unmediated Communication

We study optimal unmediated communication and provide the relevant experimental results. Instead of mediation, the players can employ unmediated communication under which the players send messages about their types to one another, after which they decide whether to coordinate on an agreement according to a public randomization device. HMS compare the mediation mechanism and the separating equilibrium of unmediated communication game, each of which maximizes the ex ante probability of peace in its respective communication environment. They show that for a subset of their parameter space, the optimal mediation yields a strictly higher chance of peace than the optimal equilibrium of unmediated communication. However, CFP test this theoretical prediction in an experiment, finding that there is no significant difference in the chance of peace across the two. While the communication protocols in our setting differ slightly from those in HMS and CFP, we can also compare the performance of P-Max or Neutral Mediator relative to unmediated communication.

We stipulate the following communication protocol. After privately learning one's own type, both players simultaneously send unverifiable messages $m_i \in \{h, l\}$ to each other. After messages are sent and received, the two players simultaneously announce "in" or "out." If they both choose "in," then the equal split is implemented; if either player chooses "out," then war takes place and the shrunk surplus is divided according to the players' types.[51] Their strategy may also depend on the realization of a public randomization device. With probability $p(m)$, the device coordinates the players on both choosing "in," and leads to the equal split.[52] With probability $1 - p(m)$, the negotiation fails and war takes place.

We restrict attention to pure-strategy separating equilibria in which players report their types truthfully and to equilibria in which probabilities $p(m)$ are symmetric across players. Let $\tilde{p}_H \equiv p(h,h)$, $\tilde{p}_M \equiv p(h,l) = p(l,h)$, and $\tilde{p}_L \equiv p(l,l)$. We calculate the optimal equilibrium of unmediated communication that maximizes the ex ante probability of peace subject to the constraints that the players communicate their types truthfully and agree to the equal split (if demanded or coordinated on by the randomization device). The following provides the equilibrium characterization.

Proposition 5. *For the baseline model with $\gamma_H > \gamma_L$, the optimal equilibrium of the unmediated communication game is characterized by $\tilde{p}_H = 1$, $\tilde{p}_M = 0$, and $\tilde{p}_L = 1$ if $\lambda < \gamma_L$ while $\tilde{p}_H = \gamma_L/\lambda$, $\tilde{p}_M = 0$, and $\tilde{p}_L = 1$ if $\gamma_L \leq \lambda < \gamma_H$.*

Proof. The optimal separating equilibrium is characterized by the following program:

$$\min_{\tilde{p}_H, \tilde{p}_M, \tilde{p}_L} q^2(1 - \tilde{p}_H) + 2q(1-q)(1 - \tilde{p}_M) + (1-q)^2(1 - \tilde{p}_L)$$

[51] To make the comparison to the mediation protocol consistent, we also constrain the two players' demands to either an equal split or walking out.

[52] In equilibrium, the players must be willing to follow the recommendation of the public randomization device with an equal split.

subject to the incentive compatibility (IC) constraints with double deviations and the "ex-post" individual rational (IR) constraints for both types. Because messages reveal types in a separating equilibrium, players must find it optimal to accept the equal split when offered (or recommended by the public randomization device). However, because $\delta\theta > 1/2$ is assumed, an H type player facing a self-reported L type opponent never finds it profitable to get a share 1/2 instead of waging war against an L type. This implies that in the optimal separating equilibrium, it must be $\tilde{p}_M = 0$ to keep in check the ex-post IR constraint for H type. Because $1/2 \geq (1-\delta)\theta$, the ex-post IR constraint for L type is always satisfied. Taking into account $\tilde{p}_M = 0$, the IC constraints with double deviations for types H and L are characterized as follows:

$$\begin{aligned} &q(\tilde{p}_H(1/2) + (1-\tilde{p}_H)\theta/2) + (1-q)\delta\theta \\ &\geq q\theta/2 + (1-q)(\tilde{p}_L \max\{1/2, \delta\theta\} + (1-\tilde{p}_L)\delta\theta); \end{aligned} \tag{H-IC*}$$

$$\begin{aligned} &q(1-\delta)\theta + (1-q)(\tilde{p}_L(1/2) + (1-\tilde{p}_L)\theta/2) \\ &\geq q(\tilde{p}_H \max\{1/2, (1-\delta)\theta\} + (1-\tilde{p}_H)(1-\delta)\theta) + (1-q)\theta/2. \end{aligned} \tag{L-IC*}$$

The maxima on the right-hand-side of both constraints takes into account the possibility of double deviations: the player being the only one to lie may deviate from the recommendation of the equal split and collect the war payoff. Because $\max\{1/2, \delta\theta\} = \delta\theta$, the constraint (H-IC*) can be rewritten as $q(\tilde{p}_H(1/2) + (1-\tilde{p}_H)\theta/2) \geq q\theta/2$, which is always satisfied for any $\tilde{p}_H \in [0,1]$. In the constraint (L-IC*), the double deviation of misreporting one's L type and then waging war if the opponent reveals to be H type is never entertained because $\max\{1/2, (1-\delta)\theta\} = 1/2$. Now setting $\tilde{p}_L = 1$ minimizes the objective function only to relax the constraint (L-IC*) without violating other constraints. Rewriting (L-IC*), we have $q(1-\delta)\theta + (1-q)(1/2) \geq q(\tilde{p}_H(1/2) + (1-\tilde{p}_H)(1-\delta)\theta) + (1-q)\theta/2$, which gives $\tilde{p}_H \leq \frac{1-q}{q}\frac{1/2-\theta/2}{1/2-(1-\delta)\theta} = \gamma_L/\lambda$. We want to set $\tilde{p}_H$ as high as possible to minimize the objective function; if $\gamma_L > \lambda$, then we set $\tilde{p}_H = 1$, and if $\gamma_L \leq \lambda$, then we set $\tilde{p}_H = \gamma_L/\lambda$. □

By the revelation principle, for any equilibrium of any communication game, we can find an equivalent incentive-feasible revelation mechanism. So mediation yields a (weakly) higher probability of peace than unmediated communication. Proposition 5 shows that the optimal equilibrium of unmediated communication, which maximizes the ex ante probability of peace, coincides with the Neutral Mediator who is associated with the lowest ex ante probability of peace among all IIE ones. Propositions 1, 3, and 5 together imply the following.

Corollary 1. *The Neutral Mediator achieves the same ex ante probability of peace as the optimal equilibrium of the unmediated communication game. All other IIE mediators (including the P-Max Mediator) achieve a strictly higher probability of peace than the optimal equilibrium of the unmediated communication game.*

We test Corollary 1 to keep in check the theoretical result of HMS and the experimental result of CFP by comparing the performance of unmediated communication (UC), mediation under the P-Max Mediator (P-Max Mediation), and that under the Neutral Mediator (Neutral Mediation).

Hypothesis 4 (UC vs. P-Max Mediation vs. Neutral Mediation)**.**

(a) P-Max Mediation yields a higher rate of agreement than UC.

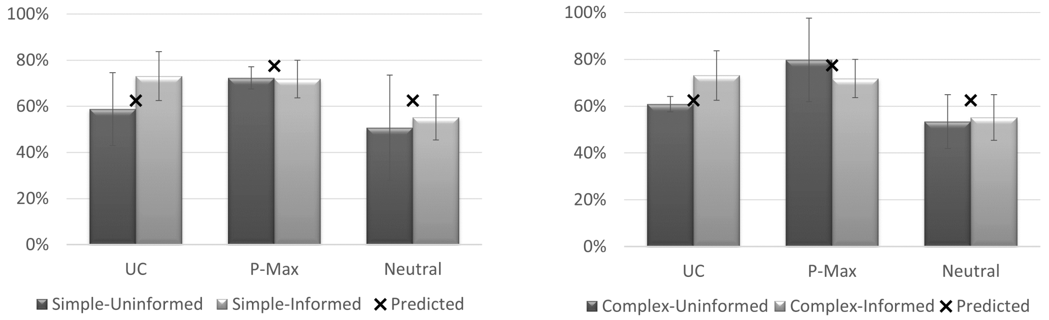
(b) P-Max Mediation yields a higher rate of agreement than Neutral Mediation.

(c) There is no significant difference in the rates of agreement between Neutral Mediation and UC.

Figure 12 illustrates the theoretically predicted rate of agreement and the observed rate of agreement across the three parts—UC, P-Max Mediation, and Neutral Mediation—for the four treatments.[53] Overall, P-Max Mediation achieves slightly higher agreement rates (72-80%) compared to both UC (59-73%) and Neutral Mediation (51-55%). These observations are qualitatively consistent with Hypothesis 4(a)-(b). The results of the Wilcoxon signed-rank tests indicate that the difference is marginally significant (one-sided, $0.05 < p$-values$<$ 0.1) for all pairwise comparisons.[54] In particular, P-Max Mediation outperforms UC and

[53] We show 95% confidence intervals calculated from standard errors clustered at the session level.

[54] When testing Hypothesis 4, we aggregate the data from the Informed and Uninformed treatments for each of the three parts. The Informed and Uninformed treatments differ only in the last part (mediator selection), so the agreement rates across the Uninformed and Informed treatments (the dark and light grey bars respectively in Figure 12) for each communication protocol (UC, P-Max, Neutral) are expected to be the same. This hypothesis is confirmed for all cases except for UC in the Simple treatment (two-sided Mann-Whitney test, p-value$= 0.0294$).



(a) Simple

(b) Complex

Figure 12: Rate of Agreement

Neutral Mediation only by a small margin. This finding aligns with CFP's experiment result that there is no significant difference in the frequency of peace between unmediated communication and mediation. Further, we cannot reject Hypothesis 4(c) as the difference is statistically insignificant (two-sided, p-value> 0.1).

Finding 6. *The rate of agreement in P-Max Mediation is marginally higher than that in Neutral Mediation or UC. The rates of agreement across Neutral Mediation and UC are not significantly different from each other.*

C.2 Experiment I: Subordinate's Best Responses

We analyze the mediation subgame (stage 2) of the mediator selection game under the subordinate's (arbitrary) interim beliefs, obtained from the prior belief probabilities $(q, 1-q)$ by updating on the basis of the principal's announced mediator. After the principal chooses and announces a mediator, the principal herself does not change her own belief on the subordinate's type, summarized by q, but the subordinate may update his belief about the principal's type being H to q', which is arbitrary. Note that here we are not theoretically examining sequential equilibria of the game per se; rather, we want to characterize the subordinate-subject's best responses when stage 2 is played given his interim belief q' in the lab. In doing so, we assume that the principal-subject report truthfully her type to be used in implementing her chosen mediator. The choice variables for the subordinate are his report about his type and his decision to go to war rather than reporting. Recall that $\theta = 0.75$, $\delta = 0.8$, and the surplus was set to 400 in the lab. Here, we only consider the case of $q = 1/4$, as we can analogously show for the case of $q = 2/5$.

First suppose that the principal announces the P-Max Mediator (as used in the lab). The expected payoffs of the H type subordinate under the belief q' (a) if he reports honestly in implementing the P-Max Mediator, (b) if he lies in implementing the P-Max Mediator, and (c) if he goes to war are, respectively, computed as follows:

$$\begin{aligned}
&\text{(a): } q'(200) + (1-q')(0.4(200) + 0.6(240)) = 200q' + 224(1-q');\\
&\text{(b): } q'(0.4(200) + 0.6(150)) + (1-q')(200) = 170q' + 200(1-q');\\
&\text{(c): } 150q' + 240(1-q').
\end{aligned}$$

The H type's payoff from (a) is always strictly greater than the payoff from (b). The payoff from (a) is greater than the payoff from (c) if and only if $q' \geq 16/66 \approx 0.2424$. The reason that the right hand side is not exactly 0.25 is because the H type's IR constraint is not binding under the P-Max Mediator with $p_M = 0.4$ used in the lab. Under the theoretical P-Max Mediator with $p_M = 5/12$, the H type's IR constraint binds, which would yield the condition to be $q' \geq 0.25$. Similarly, the expected payoffs of the L type subordinate under the belief q' are:

$$\begin{aligned}
&\text{(a): } q'(0.4(200) + 0.6(60)) + (1-q')(200) = 116q' + 200(1-q');\\
&\text{(b): } q'(200) + (1-q')(.4(200) + 0.6(150)) = 200q' + 170(1-q')\\
&\text{(c): } 60q' + 150(1-q').
\end{aligned}$$

The L type's payoff from (c) is always strictly less than the payoff from (a) or (b). The payoff from (a) is greater than the payoff from (b) if and only if $q' \leq 5/19 \approx 0.2632$. Again, the right hand side is not exactly 0.25 because the L type's IC constraint is not binding under both the theoretical P-Max Mediator and the one used in the lab. The subordinate's best response switches around the two threshold values of 0.2424 and 0.2632, which we will take them as equivalent to the prior belief of $q = 0.25$. With this caveat, the subordinate's interim beliefs $q' > 0.25$, $q' = 0.25$, and $q' < 0.25$ correspond to his inferences that, respectively, the principal is "more likely to be H than the prior," the principal is likely to be H "as the prior", and the principal is "more likely to be L than the prior.[55] We summarize below the subordinate-subject's best responses (against the principal reporting truthfully in implementing the P-Max Mediator) when stage 2 is played given his inferences (i.e., his interim belief q').

Remark 1. When the players play stage 2 after the P-Max Mediator is announced:

(i) If the subordinate's inference is "more likely to be H than the prior," then reporting truthfully is the best response for the H type while lying is the best response for the L type.

(ii) If the subordinate's inference is "same as the prior," then reporting truthfully is the best response for both types.

(iii) If the subordinate's inference is "more likely to be L than the prior," then declining the mediator is the best response for the H type while reporting truthfully is the best response for the L type.

Now suppose that the principal announces the Neutral Mediator (as used in the lab). The expected payoffs of the H type subordinate under the belief q' (a) if he reports honestly in implementing the Neutral Mediator, (b) if he lies in implementing the Neutral Mediator, and (c) if he goes to war are, respectively:

$$\begin{aligned}
&\text{(a):}\ 200q' + 240(1-q');\\
&\text{(b):}\ 150q' + 200(1-q');\\
&\text{(c):}\ 150q' + 240(1-q').
\end{aligned}$$

The H type's payoff from (a) is always strictly greater than the payoff from (b), and is strictly greater than the payoff from (c) except for when $q' = 0$ (i.e., the subordinate believes that

[55] When the prior belief is $q = 0.4$, these three cases correspond to $q' > 0.4$, $q' = 0.4$, and $q' < 0.4$.

the principal is L for sure). For the L type subordinate, the expected payoffs are:

$$\text{(a): } 60q' + 200(1 - q');$$
$$\text{(b): } 200q' + 150(1 - q');$$
$$\text{(c): } 60q' + 150(1 - q').$$

The L type's payoff from (a) is strictly greater than the payoff from (c) except for when $q' = 1$, in which case the payoff from (b) is strictly greater than either one. The payoff from (b) is strictly greater than the payoff from (c) except for when $q' = 0$, in which case the payoff from (a) is strictly greater than either one. The payoff from (a) is greater than the payoff from (b) if and only if $q' \leq 5/19 \approx 2632$. Again by taking this threshold value as equivalent to the prior belief of $q = 0.25$, we can characterize the subordinate-subject's best responses (against the principal reporting truthfully in implementing the Neutral Mediator) when stage 2 is played as follow.

Remark 2. When the players play stage 2 after the Neutral Mediator is announced:

(i) If the subordinate's inference is "more likely to be H than the prior," then reporting truthfully is the best response for the H type while lying is the best response for the L type.

(ii) If the subordinate's inference is "same as the prior," then reporting truthfully is the best response for both types.

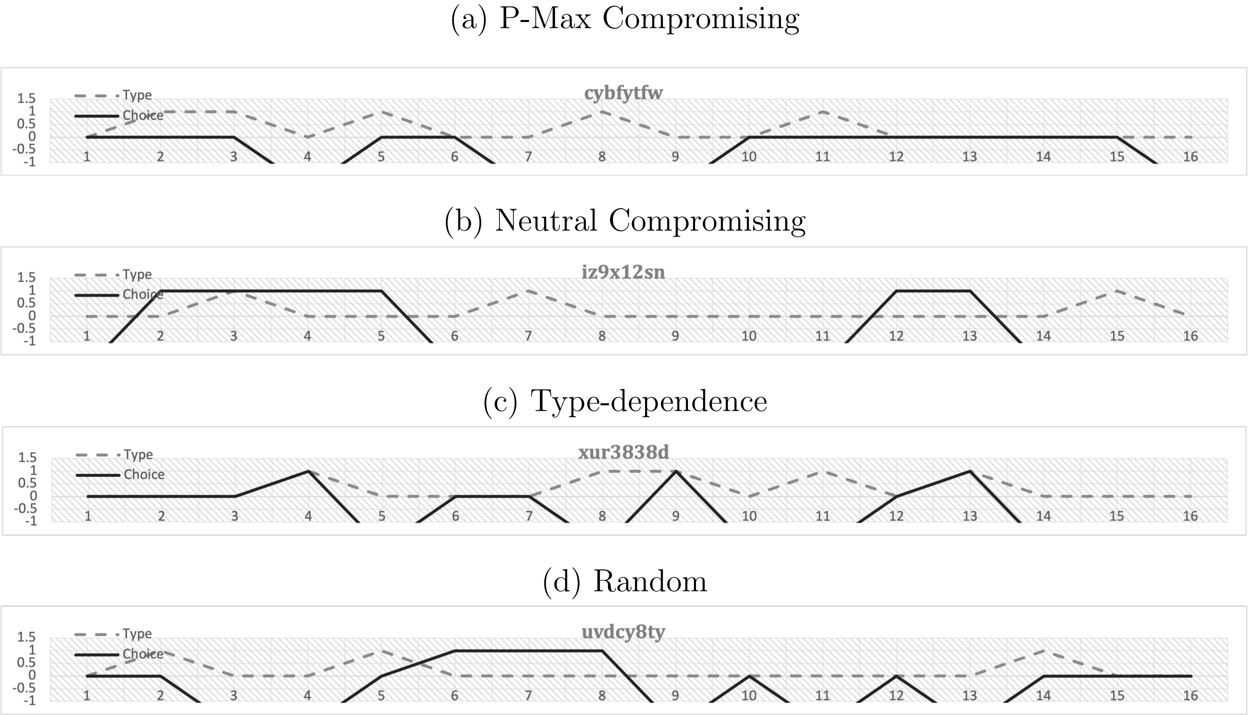
(iii) If the subordinate's inference is "more likely to be L than the prior," then reporting truthfully is the best response for both types. (If the subordinate's inference is "surely L," then both reporting truthfully and declining the mediator are best responses for the H type.)

C.3 Experiment II: Classifications of Individual Behavior

We look at the individual behavior in Experiment II more carefully. Figure 13 showcases four representative individual cases detailing their decisions as a Principal over 16 rounds.[56] In

[56] We have 90 such figures in total, which are available upon request.

each figure illustrating the choices made by an individual (with the individual ID displayed in the center-top), the values on the horizontal axis correspond to the rounds. Regarding the vertical axis values: for types (indicated by dotted lines), 1 signifies the H type, and 0 signifies the L type. Concerning choices (depicted by solid lines), 1 denotes Neutral Compromising, 0 corresponds to P-Max Compromising, -1 indicates Uncompromising, and -2 (not shown within the figures) signifies no deliberate choice as a subordinate.



(a) P-Max Compromising

(b) Neutral Compromising

(c) Type-dependence

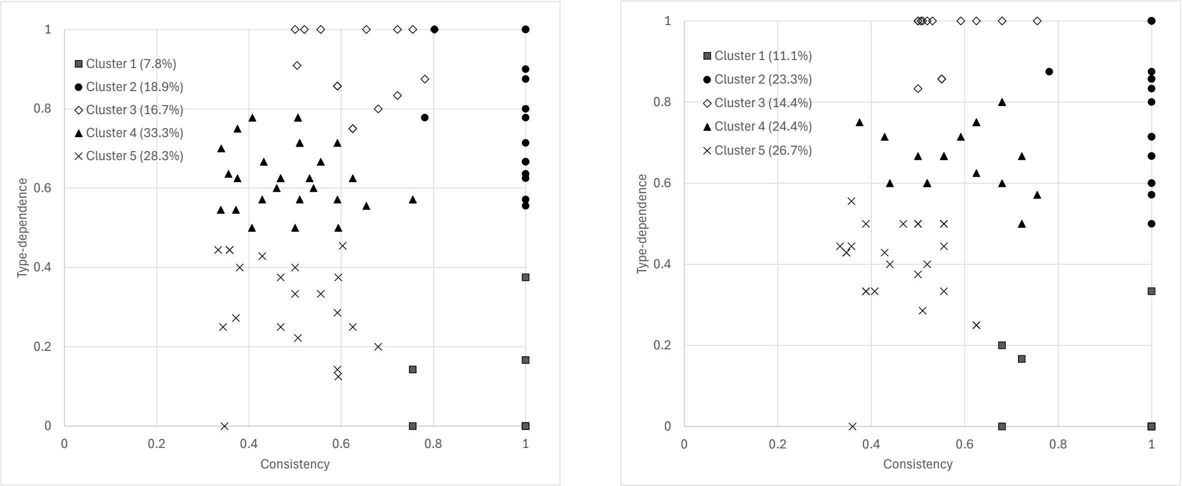
(d) Random

Note: The values in the horizontal axis represent rounds. Regarding the vertical axis: for types (dotted lines), 1 represents H type and 0 represents L type. For choices (solid lines), 1 represents Neutral compromising, 0 represents P-Max compromising, -1 represents Uncompromising, and -2 (outside the figures) represents no deliberate choice (being a subordinate).

Figure 13: Four Representative Cases

In Figure 13, Panel (a) presents the case in which the choices are consistently made for P-Max Compromising, independent of the type realizations. Panel (b) illustrates the behavior in which the choices are consistently made for Neutral Compromising, independent of the type realizations. Choices presented in Panel (c) are not consistent but perfectly dependent upon the type realizations. Panel (d) illustrates a case in which the choices are neither consistent nor type-dependent.

We consolidate all 90 individual observations using two measures: consistency and type-dependence. Consistency is quantified as the proportion of the most frequent choice, while type-dependence is calculated as the percentage of instances where the choice aligns with the type. The observations, classified by the k-means clustering method (Macqueen, 1967), into five clusters are illustrated in Figure 14. The values reported in Panel (b) of Figure 14 are calculated based on the observations excluding the first two deliberate choices.



(a) All observations

(b) Dropping the first 2 observations

Note: The horizontal axis represents the consistency of an individual's choice and the vertical axis represents the correlation between the type and the choice. The values reported in panel (a) are calculated based on all observations. The values reported in panel (b) are calculated based on the observations excluding the first two deliberate choices.

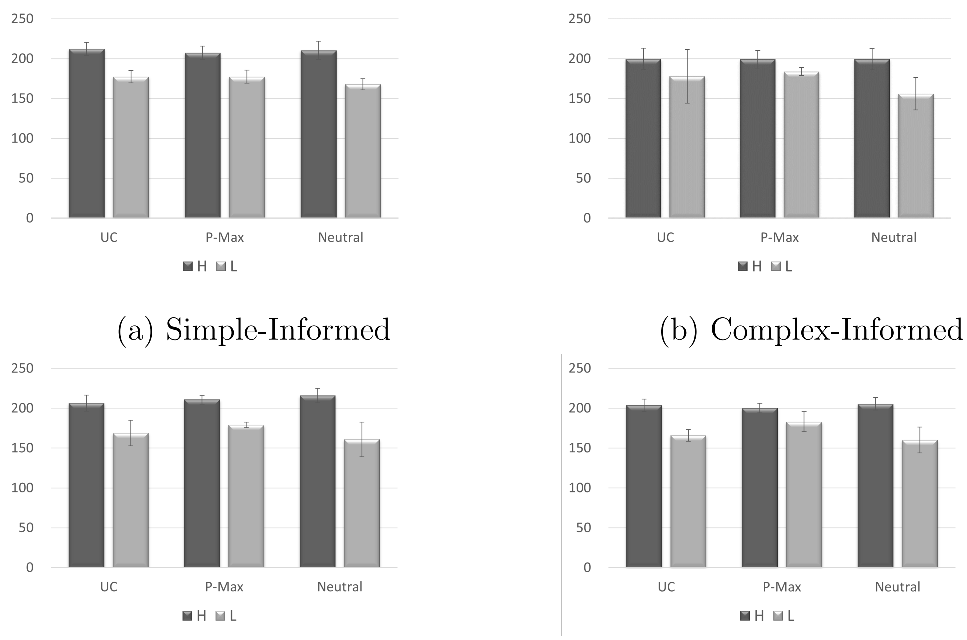
Figure 14: k-means Clustering with Five Clusters

Cluster 1 (depicted by black squares, 7.8-11.1%) comprises cases where choices exhibit consistency but lack type-dependence. The majority of observations in this cluster involve frequent selections of the Uncompromising option. Cluster 2 (illustrated by black circles, 18.9-23.3%) represents observations characterized by high consistency coupled with significant type-dependence. These instances reflect individuals who have undergone an internal deliberation process regarding the intertype compromise and have settled on one of the mediators as their inscrutable choice. Cluster 3 (shown as white diamonds, 14.4-16.7%) embodies choices with notable type-dependence but an intermediate level of consistency, indicating a failure in reaching an agreed decision between the types about the intertype compromise. Cluster 4 (depicted by black triangles, 24.4-33.3%) encompasses observations with moderate type-dependence and low consistency. This group shares similarities with Cluster 3 but exhibits more variability. Cluster 5 (represented by x-marks, 26.7-28.3%) comprises observations with both low type-dependence and consistency, indicative of random choices.

This outcome reinforces the primary finding obtained from our aggregate data analysis. It highlights that only approximately 20% of individuals engaged in the internal deliberation process of the intertype compromise to arrive at a decision regarding one of the mediators (Cluster 2). The bulk of individuals either engaged in the intertype compromise but struggled to reach an agreed compromise between the two types (Clusters 3 and 4) or did not grasp the significance of the inscrutable intertype compromise (Clusters 1 and 5).

D Additional Figures

D.1 Figures for Experiment I



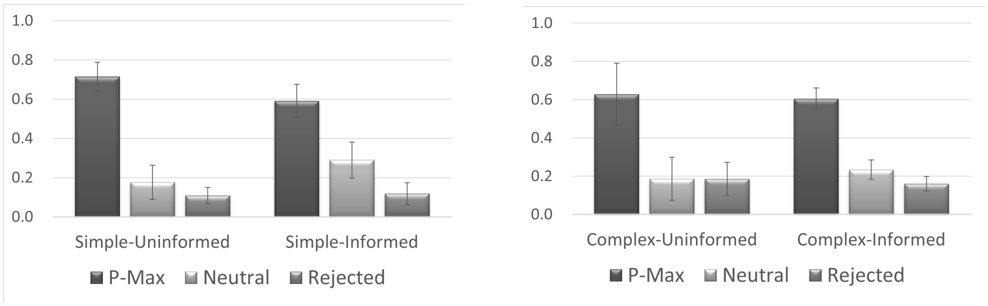
(a) Simple-Informed

(b) Complex-Informed

(c) Simple-Uninformed

(d) Complex-Uninformed

Figure 15: Average Payoff of Each Type in Parts 1–3 in Four Treatments

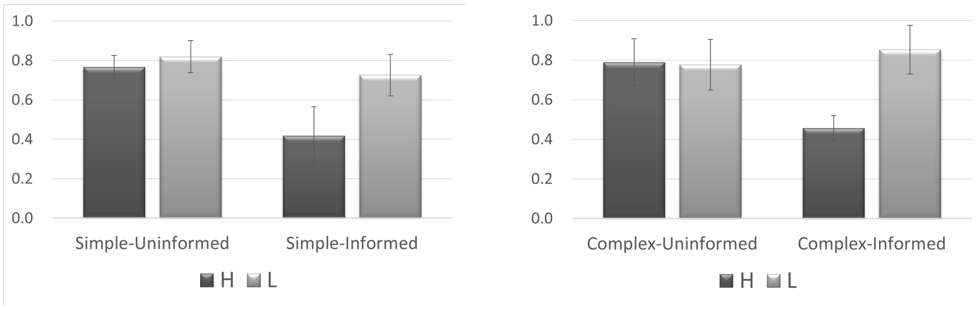


(a) Simple

(b) Complex

Note: The figure shows the proportions of mediators chosen by the principals and then played after taking into account of rejections by subordinates. So examining the data from all subjects is essentially equivalent to examining the data only from principal-subjects.

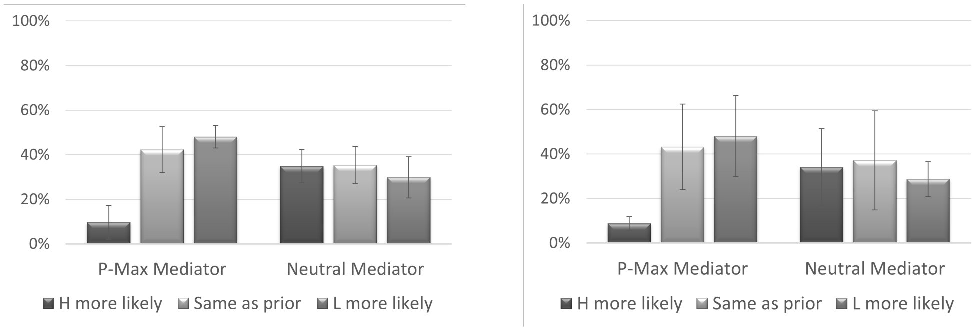
Figure 16: Proportion of Mediator Actually Played (All Subjects/Principals)



(a) Simple

(b) Complex

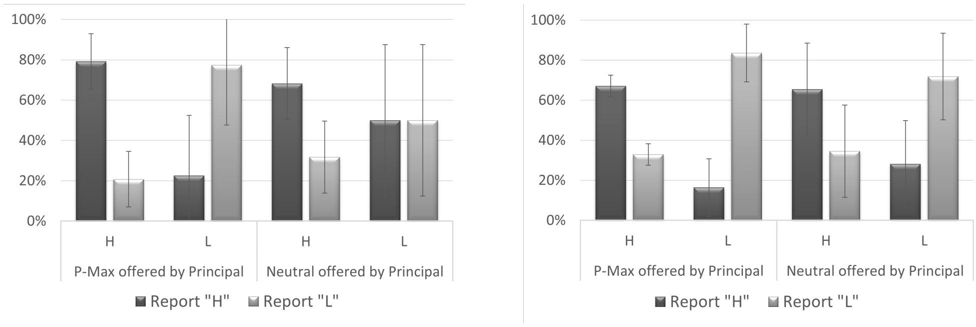
Figure 17: Proportion of P-Max Mediator Chosen By Type (Principals)



(a) Simple-Uninformed

(b) Complex-Uninformed

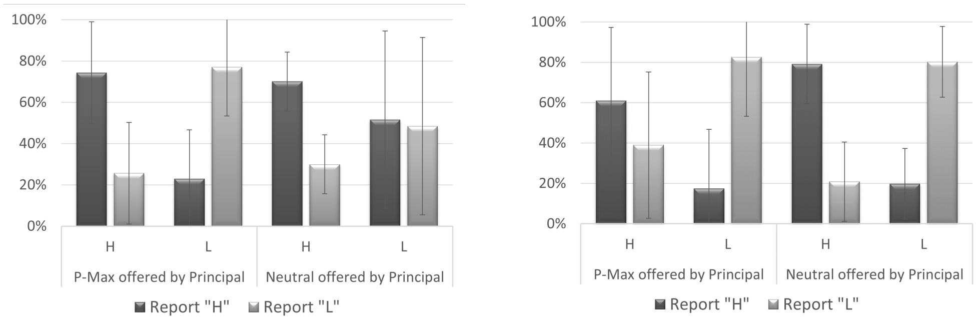
Figure 18: Subordinate's Inference Conditional on Each Mediator Chosen (Uninformed)



(a) Simple-Uninformed

(b) Complex-Uninformed

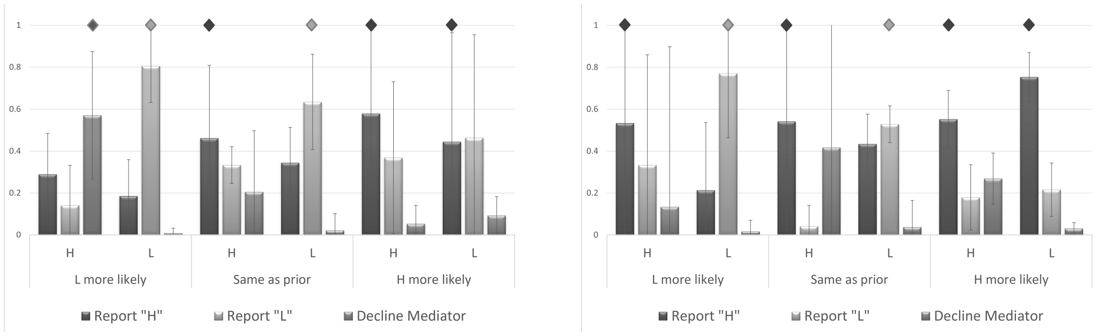
Figure 19: Principal's Strategy in Chosen Mediator by Type (Uninformed)



(a) Simple-Informed

(b) Complex-Informed

Figure 20: Principal's Strategy in Chosen Mediator by Type (Informed)

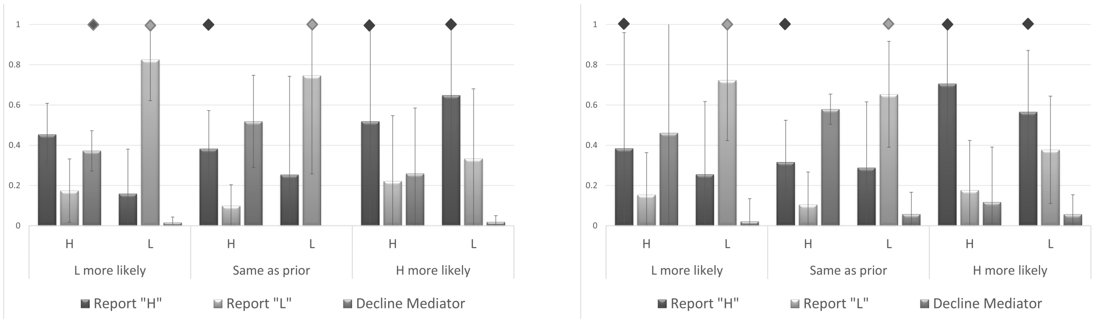


(a) When P-Max offered by Principal

(b) When Neutral offered by Principal

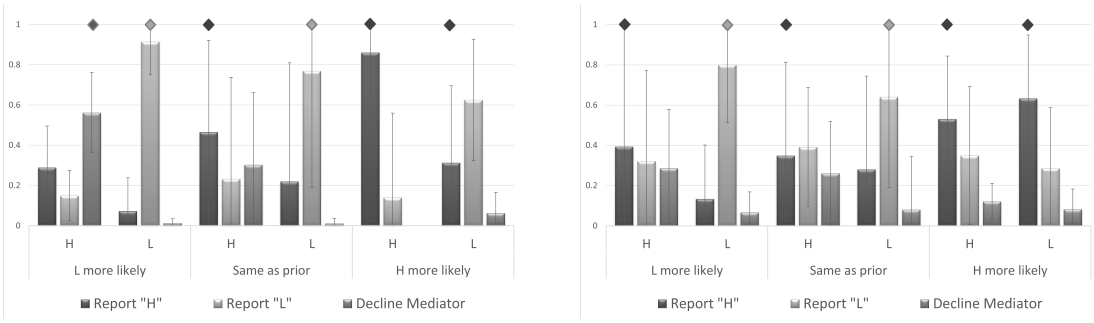
Note: The diamond shapes indicate the best response of each type to the reported beliefs in Figures 21–24.

Figure 21: Subordinate's Strategy in Chosen Mediator by Type given Inference (Simple-Informed)



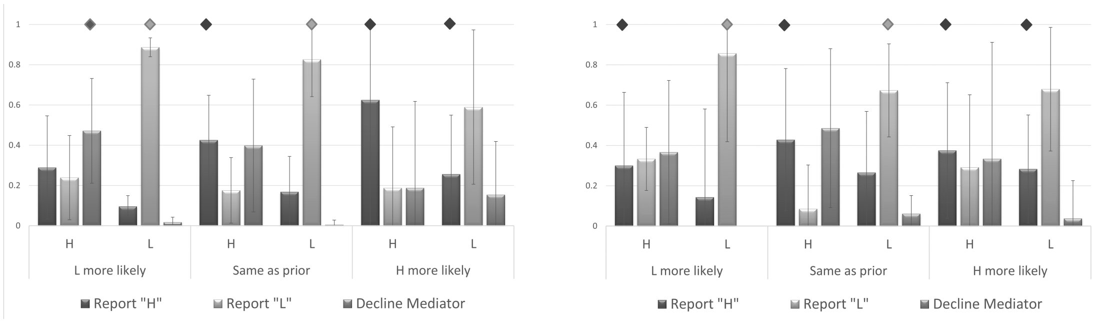
(a) When P-Max offered by Principal (b) When Neutral offered by Principal

Figure 22: Subordinate's Strategy in Chosen Mediator by Type given Inference (Simple-Uninformed)



(a) When P-Max offered by Principal (b) When Neutral offered by Principal

Figure 23: Subordinate's Strategy in Chosen Mediator by Type given Inference (Complex-Informed)

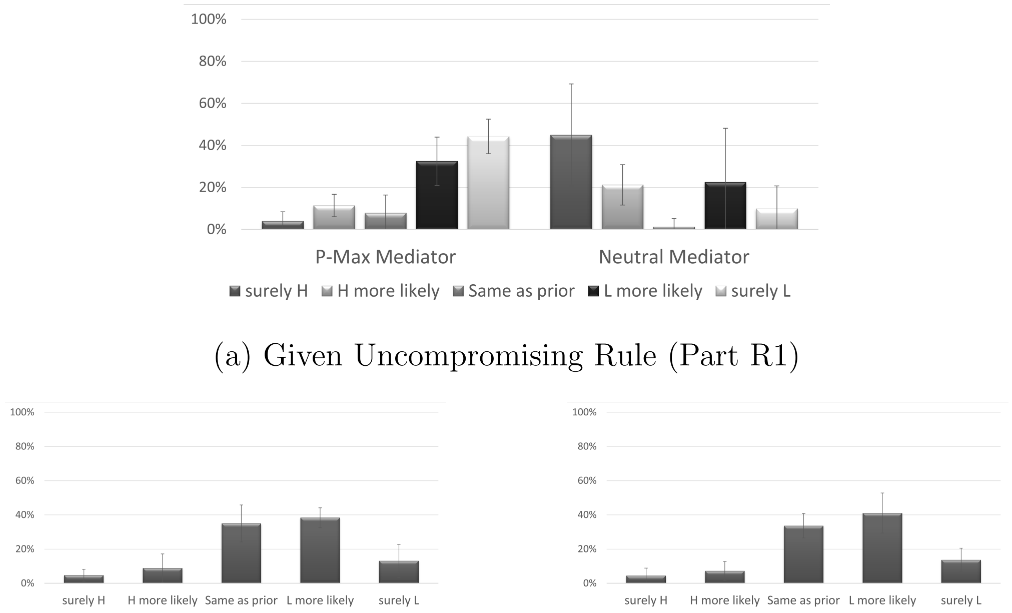


(a) When P-Max offered by Principal (b) When Neutral offered by Principal

Figure 24: Subordinate's Strategy in Chosen Mediator by Type given Inference (Complex-Uninformed)

D.2 Figure for Experiment II (Parts R1-R3)

Figure 25 shows the frequencies of five inferences by subordinates conditional on either P-Max or Neutral Mediator chosen by the principal according to each given mediator-selection rule. Given the Uncompromising rule (Part R1), most subjects make "L" inferences after observing the P-Max Mediator and "H" inferences after observing Neutral Mediator. Given either the P-Max or Neutral Compromising rule (Parts R2 and R3), most subjects make "same as prior" or "L more likely" inferences (73.3% under P-Max Compromising and 74.7% under Neutral Compromising). We conjecture that some subjects interpret the "L more likely" inference to be equivalent to the "same as prior" inference because the prior probability of L type is already 3/4.



(a) Given Uncompromising Rule (Part R1)

(b) Given P-Max Compromising Rule (Part R2)

(c) Given Neutral Compromising Rule (Part R3)

Figure 25: Subordinate's Inference Conditional on Chosen Mediator given Selection Rule