Toward an Understanding of Optimal Mediation Choice*

Jin Yeub Kim† Wooyoung Lim‡

March 3, 2026

Abstract

Mediation is a key strategic instrument for managing conflicts in bargaining scenarios with incomplete information. This paper reports the first systematic laboratory investigation into the informed principal problem concerning mediator selection. The theory of neutral optimum predicts that, in our environment, the informed principal's most reasonable choice is not the mediator that maximizes the ex-ante probability of peace; rather, the one preferred by the stronger type alone constitutes a credibly justifiable compromise between the conflicting interests of different types. We find that subjects do not choose the neutral mediator more often than the peace-maximizing one. Different principal types recognize the need for inscrutable selection and form intertype compromises, and they systematically view the peace-maximizing mediator as the more compelling compromise. The strategic reasoning underlying the neutral optimum fails to materialize in the lab.

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, Indiana University Microeconomics seminar, Korea University Economics Department seminar, the 2023 SAET Conference in Paris, 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 Alex Kim, 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

‡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 forms of third-party intervention for dispute resolution across a wide range of conflicts, from labor-management and legal disputes to international conflicts.[1] 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 analyses of mediation, often adopting a mechanism design approach, treat mediators as exogenous and assume that their objective is to minimize 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. Yet, in practice, disputants often choose a mediator themselves from among many available mediators who may differ in their effectiveness in bringing peace.[2]

This raises a fundamental question: What kinds of mediators should we expect disputants to choose? The answer is not obvious, particularly 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 she most prefers. On the other hand, her mediator choice may reveal private information she would rather conceal. In this paper, we study how such a dilemma is resolved, that is, how concerns about mediator selection potentially leaking private information shape the principal's choice of mediator.

The analysis of mechanism selection by an informed principal offers theoretical insights. In his seminal contribution, Myerson (1983) develops a theory of inscrutable mechanism selection and introduces several solution concepts that satisfy *inscrutability*: the informed principal should choose a mechanism that is a reasonable selection for all of her possible types, so that the selection itself conveys no information. If different types of the principal

[1]The use of alternative dispute resolution (ADR) techniques, such as arbitration and mediation, in civil trials have increased substantially 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 authorize some form of ADR, two thirds of which involve mediation. For the types of legal cases where mediation may 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 February 15, 2026). For incidences of mediation in international conflicts, see Bercovitch and Gartner (2009) and Frazier and Dixon (2006).

[2]The American Arbitration Association offers mediation to parties in various disputes in industries and fields, and provides a roster of mediators from which parties may choose (see `https://www.aaamediation.org/find-a-mediator`, accessed February 15, 2026). Individual mediator profiles report information on experience, style, and procedural preferences, all of which may impact mediation success rates. Wilkenfeld et al. (2003) document, using historical data and experimental approach, that mediation effectiveness varies with mediator style in international crises.

prefer different mediators, then she must select the one that can be perceived as a reasonable *intertype compromise* between the conflicting preferences of those types. A central concept that characterizes such a compromise is the concept of *neutral optimum*.[3]

We take the informed principal's mechanism selection problem to the lab. To our knowledge, this is the first experimental study of the informed principal problem. To ensure analytical tractability and salient incentives, we use a simple conflict environment with a two-type information structure and two alternative mediators implemented by computer algorithms. Our primary objective is to test the prediction of the neutral optimum and to examine subjects' behavior with respect to 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 both the model and the experiment, two players compete over a fixed surplus. They can jointly opt for an agreement to split the surplus equally, or either player can inflict disagreement which shrinks the surplus. Each player has a private type, either high or low, that determines players' disagreement payoffs. The high type prefers agreement over disagreement only when facing the same type, whereas the low type always prefers agreement regardless of the opponent's type. In this setting, players can use mediation, under which they confidentially send messages about their types to a mediator with a commonly known algorithm; then the mediator issues a message-dependent recommendation of agreement or disagreement.[4]

We consider situations in which an informed principal chooses between two 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 (described below), and can also be justified as an inscrutable intertype compromise. The concept of neutral optimum (Myerson, 1983) predicts that, in our setting, pooling on the neutral mediator, which is preferred by the high type, constitutes the most compelling intertype compromise. By contrast, the peace-maximizing mediator corresponds to the one that maximizes the ex ante probability of agreement and is preferred by the low type. It is the unique ex ante incentive efficient mediator in our setting, thus the natural choice for an uninformed principal who chooses prior to learning her type.

[3]The neutral optimum is an axiomatically founded solution concept that cannot be blocked with any reasonable theory of blocking (Myerson, 1983). See Section 3.3 for more details.

[4]Hörner, Morelli and Squintani (2015) describe a recommendation of disagreement as a mediation failure leading to players fighting a war. Casella, Friedman and Perez Archila (2025) describe it as the mediator "walking out," which is also the terminology we adopt. The main distinction in our setting is that only a single agreement outcome (an equal split) is possible. Further discussion of our modeling choice is provided in Section 2.1 and in online Appendix C.1.

[5]The theory of efficient mechanisms (Holmström and Myerson, 1983) identifies both as plausible selections for an informed principal.

We conducted two experiments in sequence. The first experiment was designed to collect observational data to test the theory of neutral optimum. After finding that the theory did not work in the lab, we conducted a second experiment to identify the source of the failure and to examine whether players made inscrutable selections.

As the main treatments of our first experiment, we considered two versions of the principal's mediator selection game that differed in the information structure at the time of selection. In the *uninformed* mediator selection game, the principal chose a mediator before either player learned its private type. In the *informed* mediator selection game, after the players learned their private types, the principal chose a mediator. After the chosen mediator was announced in both versions, the other player (subordinate) was asked to report his belief about the principal's type. Finally, each player confidentially reported its type to the chosen mediator; at this stage, the subordinate could decline the mediator, thereby inflicting disagreement, instead of submitting a report.

We find that the peace-maximizing mediator is chosen significantly more often than the neutral mediator in the uninformed selection game, consistent with theory. However, in the informed selection game, the neutral mediator is not chosen more often than the peace-maximizing mediator, contrary to the neutral optimum prediction. More importantly, the two principal types do not select the same mediator; and low types choose the peace-maximizing mediator over the neutral mediator more often than high types choose the neutral mediator over the peace-maximizing mediator. This asymmetric, type-dependent pattern may reflect a failure to understand the benefit of inscrutable selection. It may also arise even if both types recognize the need for inscrutable selection and form an intertype compromise, yet disagree on how that compromise should be resolved.

To distinguish between these explanations, we conducted a second experiment. In the main part of this experiment, after learning their private types, principals were asked to choose among three mediator-selection rules that specify both their realized type's mediator choice and the mediator that would be chosen by their unrealized type: (1) choose the mediator preferred by the realized type and the other under the unrealized type (Uncompromising); (2) choose the peace-maximizing mediator under both types (P-Max Compromising); or (3) choose the neutral mediator under both types (Neutral Compromising). This design allows us to examine whether informed principals choose inscrutably, and if so, to elicit their views on how their unrealized type should behave when forming an intertype compromise.[6] Importantly, although subjects made their mediator choices after observing their realized type, our

[6]Unlike in the strategy method (Selten, 1967), principals made their choices after observing their realized type, so the mediator specified for the unrealized type could not be implemented directly. As a result, subjects made their decisions under interim incentives rather than from an ex ante perspective.

bonus-payment scheme rendered their unrealized type's choice payoff-relevant. The principal's mediator choice was automatically assigned based on the selected rule and her realized type, and the resulting mediator was announced. The subordinate observed only the announced mediator, and the remainder of the process proceeded as in the first experiment.

Our experimental evidence indicates that *the theory of inscrutable mechanism selection is supported, whereas the concept of neutral optimum is not.* We find that subjects rarely choose the Uncompromising option and instead choose one of the two Compromising options significantly more often. This pattern suggests that subjects recognize the need for inscrutable selection and seek an intertype compromise between the objectives of their true type and their unrealized type. We further find that both types choose the P-Max Compromising option significantly more often than the Neutral Compromising option. This tendency suggests that *both types appear to regard the peace-maximizing mediator as the more compelling intertype compromise.*

These findings in the second experiment raise a question: if subjects choose inscrutably, why does the average pattern of mediator choices in the first experiment exhibit asymmetric type dependence? The data on mediator-selection rule choices by type clarify that the empirically observed type-dependent mediator choices do not directly reflect a breakdown of inscrutable selection. Rather, they arise from two behavioral sources. First, a fraction of principals of both types fail to choose pooling rules, contributing symmetrically to divergence across types. Second, and more importantly, among the majority who choose pooling rules, disagreement persists over how the intertype compromise should be resolved. In particular, a nontrivial fraction of high types prefer pooling on the neutral mediator, diverging from the rest of the population. When these rule choices are mapped into the mediators actually implemented, the resulting pattern closely reproduces the asymmetric type-dependence observed in the first experiment.

The neutral optimum in our setting relies on an asymmetric strategic structure. In theory, if mediator choice reveals the principal to be the low type, a high type subordinate can credibly threaten disagreement, making the peace-maximizing mediator costly for low type principals. By contrast, if the principal is revealed to be the high type, a low type subordinate cannot mount a comparable threat. For the neutral mediator to emerge as an intertype compromise, principals must internalize this asymmetry ex ante and anticipate that choosing the peace-maximizing mediator would trigger sufficiently strong strategic discipline from high type subordinates.

Evidence from the mediation stage sheds light on why this logic fails in the lab. Although subordinates update their beliefs in response to mediator choice, the behavioral implications of these (unincentivized) updates are limited. High type subordinates do not decline the

peace-maximizing mediator often enough to discipline low type principals, while low type subordinates' behavior under the neutral mediator reduces its relative attractiveness for high type principals. Thus, the asymmetric credible-threat structure that underpins the neutral optimum does not materialize. Principals instead converge on an inscrutable compromise tilted toward the peace-maximizing mediator, and the neutral optimum fails to emerge.

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 subsequently developed taking a different approach by Maskin and Tirole (1990, 1992). A few studies examine the problem in private value environments (Cella, 2008; Maskin and Tirole, 1990; Mylovanov and Tröger, 2012, 2014), while several others analyze it in common value environments (Balkenborg and Makris, 2015; Dosis, 2022; Kim, 2017; Koessler and Skreta, 2016, 2019; Myerson, 1983; Maskin and Tirole, 1992; Nishimura, 2022; Severinov, 2008; Skreta, 2011). We consider a common value environment and use the concept of neutral optimum to characterize a possible mediator choice. Our contribution is to provide a direct 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 study to ours is Casella, Friedman and Perez Archila (2025), 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; however, Casella, Friedman and Perez Archila (2025) find no such improvement in their experiment. In our first experimental design, subjects also play an unmediated communication game. Our evidence is mixed: peace-maximizing mediation yields a higher frequency of agreement than unmediated communication, whereas neutral mediation yields a lower frequency. While the theoretical literature on the problem of informed principal's mechanism selection and on the effectiveness of mediation is quite large, experimental contributions remain limited. To our knowledge, this is the first paper to experimentally study informed principal's mechanism selection in a mediation setting, taking a step toward understanding disputants' mediation choice in practically relevant conflicts.

The paper is organized as follows. In Section 2, we describe the baseline environment, the mediator selection game and the model parameterization. In Section 3, we provide the theoretical predictions and present our hypotheses. In Section 4, we describe the experimental design and procedure. In Sections 5 and 6, we report our experimental findings. In

Section 7, we conclude and discuss possible extensions. Appendix A provides nonparametric test results; and Appendix B reports supplementary data and analyses that are not included in the main text but are referred to therein. Theoretical discussions and characterizations, along with additional analyses of non-essential data from the main experiments and data from an auxiliary experiment, are contained in the Online Appendix. The *Experimental Instructions* (url-linked) are available on the authors' websites.

2 The Model

2.1 Environment

We consider a simplified version of a standard conflict model studied in Hörner, Morelli and Squintani (2015) (HMS) and used in Casella, Friedman and Perez Archila (2025) (CFP).

Two players (1 and 2) dispute a surplus of size one. 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. War shrinks the value of the surplus to $\theta < 1$, which is divided according to the players' types: when both players are of the same type, each player's war payoff is $\theta/2$; when types differ, the H type player receives a share $\delta > 1/2$, yielding payoff $\delta\theta > 1/2$, while the L type player receives $(1-\delta)\theta$. War occurs unless both players agree to an equal split, in which case the split is implemented and each player receives one half regardless of type. We focus on cases where $q\theta/2 + (1-q)\delta\theta > 1/2$, otherwise the equal split is preferable to war for both types.

In this setting, the players can communicate through mediation. We set up the mediation game as a direct-revelation mechanism of the following form: After being informed of their own type, each player sends a confidential message $m_i \in \{h, l\}$ to a mediator. Given reports $m = (m_1, m_2)$, the mediator publicly recommends an equal split $(1/2, 1/2)$ with probability $p(m)$ or war with probability $1 - p(m)$.[7] In the lab, a recommendation of war is phrased as the mediator "walking out," resulting in players fighting a war, as used by CFP. We focus on symmetric recommendation probabilities, so we let $p_H \equiv p(h,h)$, $p_M \equiv p(h,l) = p(l,h)$, and $p_L \equiv p(l,l)$. The mediator commits to its mechanism, or mediation plan, (p_H, p_M, p_L); we use the terms "mediator" and "mechanism" interchangeably.

[7] Our environment and mediation protocol differ from those in HMS in that the only possible agreement is an equal split. This restriction simplifies the representation of mediators, as each can be characterized solely by its recommendation probabilities, and it streamlines the mediation protocol by obviating the need for a separate stage of accepting or rejecting recommendations. Because our focus is on the initial mediator selection rather than on the mediator's recommendation strategy or the players' strategic behavior during mediation, this simplification is without loss for our purposes. See online Appendix C.1 for a more detailed discussion of our modeling choice.

2.2 Mediator Selection Game

We study situations in which the players can choose among mediators that differ in their mediation plans. We call the player choosing a mediator the *principal* and the other player the *subordinate*. The problem of mediator selection belongs to the class of informed principal problems; our main objective is to experimentally investigate the theory of mechanism selection by an informed principal.

We consider two information structures: In *informed mediator selection*, the principal chooses a mediator after each player learns its own type. In *uninformed mediator selection*, the principal chooses a mediator before any player learns its type.

In the informed mediator selection game, the timing is as follows.

1. Each player first learns its own type; then the informed principal selects and announces a mediator.
2. Each player confidentially reports its type to the announced mediator, in which case the mediator's recommendation is implemented. 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. Because war can be initiated unilaterally, the subordinate can trigger war whenever it becomes profitable given his inference after the announcement. In the lab, this option is phrased as *declining* the announced mediator; otherwise, the subordinate participates in mediation by sending a report in Stage 2.[8] By contrast, the principal cannot infer anything about the subordinate's type from her own announcement, and announcing a mediator can itself be interpreted as a decision to participate in mediation; therefore, the principal's option to trigger war in Stage 2 can be assumed away.[9]

In the uninformed mediator selection game, an *uninformed* principal first selects and announces a mediator; then each player learns one's own type in Stage 1, after which the game proceeds to Stage 2 in the same manner. We assume that the principal commits to her announced mediator; allowing the principal to change her choice after types are realized only complicates the analysis both in theory and in the lab.

[8] If Stage 2 were separated so that the subordinate first decides whether to go to war or participate in the mediation, and then decides his report conditional on participation, the principal may also be able to infer information about the subordinate's type from the participation decision. This only adds another layer of the information leakage problem, which requires expanding the principal's viable actions, complicating both the theoretical analysis and the experimental design without providing commensurate insights. Further, only the subordinate is allowed to decline in Stage 2, as only his participation constraint may be affected by belief updating following the mediator choice. See online Appendix C.2 for more details.

[9] One could allow the principal to forgo mediator selection in Stage 1 and immediately go to war. However, this extension detracts from our focus, which is on how the principal selects among mediators when such a choice is available.

We delineate the pool of different possible mediators as the set of *interim incentive efficient* (IIE) mechanisms (Holmström and Myerson, 1983). A mechanism is feasible if it is incentive compatible, so that every player wishes to report its 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. In our environment, there are infinitely many such mechanisms (mediators).[10]

2.3 Experimental Parameterization

Throughout the experiments, we fix $\theta = 0.75$ and $\delta = 0.8$. These parameters are chosen so that the two types' different preferences over outcomes are salient. In the lab, the equal split is described as agreement and war as disagreement. To avoid decimals, the size of the surplus is set to 400. Given these parameters, Table 1 displays the players' payoffs as a function of the outcome and the type profile; in each cell, the first entry indicates player 1's payoff and the second entry player 2's payoff.

Table 1: Payoffs

	(H,H)	(H,L)	(L,H)	(L,L)
Agreement	200, 200	200, 200	200, 200	200, 200
Disagreement	150, 150	240, 60	60, 240	150, 150

We consider two values of the prior probability of the H type: $q = 1/4$ and $q = 2/5$.[11] For each q, two mediators are provided as available choices, whose mediation plans are shown in Table 2. Each mediator is labeled according to its percentage value of p_M. The table also reports the ex ante probability of agreement under each mediator.

Table 2: Two Mediators' Mediation Plans

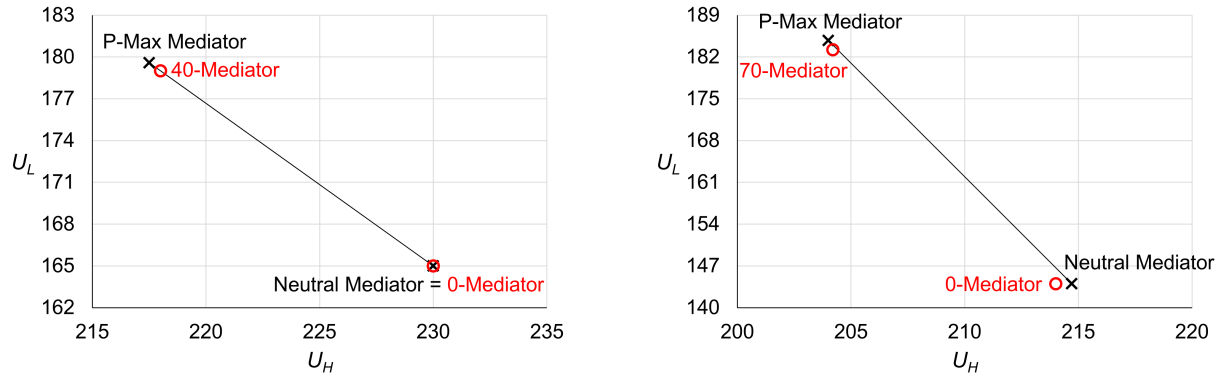
		p_H	p_M	p_L	$P(peace)$
$q = 1/4$	40-Mediator	1	.40	1	77.5%
	0-Mediator	1	0	1	62.5%
$q = 2/5$	70-Mediator	.85	.70	1	83.2%
	0-Mediator	.50	0	1	44.0%

[10] We treat interim efficiency as a maintained assumption rather than deriving it from the mediation-selection process. In online Appendix C.3, we fully characterize the set of IIE mediators in our environment and discuss an alternative approach to mediator selection following Maskin and Tirole (1992), who characterize the set of equilibrium allocations in their mechanism selection game without imposing interim efficiency.

[11] CFP use $\theta = 0.7$ and $\delta = 1$ as their experimental parameters, and consider $q = 1/3$ and $q = 1/2$.

The two mediators used in the lab are chosen to represent two distinct theoretical benchmarks among the many IIE mediators in our environment. The first benchmark is associated with the highest ex ante probability of agreement; we refer to this mediator as the *P-Max Mediator*. The second benchmark corresponds to the neutral optimum in Myerson's (1983) theory of inscrutable mechanism selection (explained in Section 3.3); we refer to this mediator as the *Neutral Mediator*. In the lab, the P-Max Mediator is represented by the 40-Mediator when $q = 1/4$ and by the 70-Mediator when $q = 2/5$, while the Neutral Mediator is represented by the 0-Mediator for both values of q.

Figure 1 depicts the expected-payoff pairs for the H and L types in all IIE mediators, with the theoretical benchmarks indicated by $\times$ markers. The mediators used in the lab, indicated by $\circ$ markers, are close approximations to the corresponding theoretical benchmarks.



(a) For $q = 1/4$ (b) For $q = 2/5$

Figure 1: The Expected Payoffs of the H and L Types, (U_H, U_L), in IIE Mediators

Note: The line depicts the frontier of expected payoffs generated by the set of IIE mediators. The $\times$ markers indicate the expected payoffs under the P-Max and Neutral Mediators. The $\circ$ markers indicate the expected payoffs under the approximations, 40-/70-Mediator and 0-Mediator, used in the lab.

We adopt the approximations for two main reasons. First, we simplify the presentation of the mediation plans to facilitate subjects' understanding. The P-Max Mediator features $p_M = 5/12 \approx .4167$ when $q = 1/4$, and $(p_H, p_M) = (630/721, 75/103) \approx (.8738, .7282)$ when $q = 2/5$, while the Neutral Mediator has $p_H = 15/28 \approx .5357$ when $q = 2/5$. Using integer percentage values, as specified in Table 2, instead of these exact theoretical values makes the mediation plans easier to compare without altering their qualitative properties. Second, we design the approximations to avoid binding incentive constraints, thereby making incentives more salient in the experiment. Under the P-Max Mediator, the H type's participation constraint binds for both values of q. Using $p_M = .40$ when $q = 1/4$ and $p_M = .70$ when $q = 2/5$ instead renders this constraint slack. When $q = 2/5$, the L type's incentive compatibility constraint binds under both the P-Max and Neutral Mediators; adjusting p_H to .85 and .50,

respectively, relaxes this constraint.[12] With all constraints slack, the 40-Mediator for $q = 1/4$ lies on the IIE frontier, while the 70- and 0-Mediators for $q = 2/5$ fall slightly below but remain close to the frontier. Because our primary interest is in subjects' choices between the two extremes along the IIE frontier, rather than in exact implementation of the frontier itself, we view the lab mediators as reasonable approximations to their theoretical counterparts.

Restricting attention to two options simplifies subjects' decision problems without complicating data with random choices.[13] When $q = 1/4$, the two mediators differ only in the value of p_M, making comparisons particularly transparent; but subjects may perceive the H type as too unlikely. When $q = 2/5$, the H type is more likely, but the mediation plans are more complex. Using both values of q allows us to see whether prior probabilities or the complexity of mediation plans affects subject behavior, although the two effects cannot be separately identified in our design because the mediation plans themselves depend on q.

3 Theoretical Predictions and Hypotheses

We derive theoretical predictions for both uninformed and informed mediator selection over the full set of IIE mediators and map them into testable hypotheses for the two mediators used in the lab.

3.1 Uninformed Mediator Selection

If the mediator is selected before types are realized, the appropriate efficiency concept is ex ante incentive efficiency in the sense of Holmström and Myerson (1983). Accordingly, in Stage 1 of the uninformed mediator selection game, the principal evaluates mediators based on her ex ante expected payoff. Among all IIE mediators, the P-Max Mediator maximizes the principal's ex ante expected payoff. Given the requirement of ex ante incentive efficiency, the uninformed mediator selection game has an equilibrium supporting the P-Max Mediator, whereas no other IIE mediators can be sustained in equilibrium. Hence, under this weak implementation concept, the P-Max Mediator is the most reasonable choice for an uninformed principal. This leads to the following prediction.

Hypothesis 1 (Uninformed Mediator Selection)**.** *Uninformed principals choose the P-Max Mediator over the Neutral Mediator.*

[12] When $q = 2/5$, the L-type incentive compatibility requires $p_H \leq 241/280 \approx .8607$ given $p_M = .7$, and $p_H \leq 15/28 \approx .5357$ given $p_M = 0$. Setting p_H as high as possible maximizes the H type's interim payoff without affecting the L type's interim payoff, highlighting a trade-off between relaxing the L-type incentive constraints and exact interim incentive efficiency.

[13] We do not expect adding more options would affect the qualitative insights of the experimental results.

3.2 Informed Mediator Selection: Pooling and Inscrutability

When mediator selection occurs after the principal learns her type, she evaluates mediators based on her interim expected payoff given her realized type. In our environment, different types strictly prefer different IIE mediators, so a naive intuition would suggest type-contingent mediator choices. Such behavior, however, cannot be sustained in equilibrium with honest participation. If the two types were to select different mediators, the subordinate would infer the principal's type from the mediator choice itself. Once this inference is made, the selected mediator becomes infeasible. As a result, no separating equilibrium with honest participation exists.[14]

By contrast, any IIE mediator that is selected with probability one regardless of the principal's type, and is followed by participation and truthful reporting, can be supported as a sequential equilibrium of the informed mediator selection game.[15] This yields a sharp behavioral prediction: *both types of informed principals must pool on the same IIE mediator*. Importantly, the prediction is purely equilibrium-based and does not depend on which mediator is chosen, provided that the choice is independent of the principal's type. It is not a direct consequence of 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 analytical device for focusing on pooling equilibria rather than generating any specific behavior prediction.

When the principal's different types prefer different IIE mediators, sustaining a type-independent choice requires more than equilibrium reasoning alone. Myerson (1983) develops a theory of inscrutable mechanism selection that addresses precisely this problem. An *inscrutable selection* captures the idea that any predicted mediator must be reasonable for all possible types of the principal. Thus, the chosen mediator must embody some form of compromise between the conflicting objectives of the principal's different possible types, rather than merely satisfy the pooling condition. Myerson (1983) provides several notions of how the informed principal should make such an *inscrutable intertype compromise*.

A key feature of this notion is that it requires an implicit thought process on the part of the principal before making a mediator choice. Although the principal already knows her realized type at the time of selection, she must internally contemplate how the mediator choice would be evaluated if her type were different. A mediator is inscrutable only if the principal can coherently assign it as the choice of *both* her realized type and her unrealized

[14] See online Appendix C.4 for a formal proof.

[15] Because pooling selection preserves prior beliefs and any IIE mediator is incentive compatible and individual rational under those beliefs, truthful participation following type-independent selection constitutes a sequential equilibrium.

alternative type. Inscrutability therefore requires not only pooling, but also that the pooled choice be justifiable as a reasonable intertype compromise from the perspective of each type.

These considerations motivate the following testable implications.

Hypothesis 2 (Informed Mediator Selection: Inscrutable Intertype Compromise)**.**

(a) Both H and L types of informed principals choose the same mediator.

(b) Informed principals believe that their unrealized type should choose the same mediator as their realized type.

Part (a) captures the equilibrium logic of pooling: the informed principal's mediator choice must be type-independent. Rejecting this part of the hypothesis, however, should not be interpreted as evidence against inscrutable selection. Any observed type-dependent behavior may arise either from a lack of understanding of inscrutable selection or from divergent views on the intertype compromise. Part (b) examines the stronger notion of inscrutability, asking whether principals consciously form an intertype compromise rather than relying on type-specific reasoning that merely leads to the same mediator. The two parts together allow us to distinguish a failure of inscrutable selection itself from disagreement over which mediator constitutes the appropriate intertype compromise.

3.3 Informed Mediator Selection: Neutral Optimum

Pooling equilibria can be viewed as satisfying a weak form of intertype compromise: a mediator is a reasonable selection for all types insofar as it is sequentially rational for each type to select it and participate honestly. However, as multiple IIE mediators can be sustained as pooling equilibria, equilibrium reasoning alone does not deliver a sharper prediction about which mediator should prevail. Standard equilibrium refinements likewise remain silent on how conflicts across the principal's possible types ought to be resolved.

To address this indeterminacy, Myerson (1983) introduces the concept of neutral optimum, which identifies the mediator that provides the most compelling compromise among the principal's types.[16] The central idea is to rule out mediators that can be blocked; that is, those for which some type of the principal could credibly argue, once the informational implications of mediator choice are taken into account, that an alternative arrangement would be strictly preferable. A mediator is a neutral optimum if it survives all such blocking arguments. In this sense, the neutral optimum represents the strongest form of reasonable

[16]For the other notions of intertype compromise developed in Myerson (1983), no safe and undominated mechanism exists in our environment, while core mechanisms and expectational equilibria coincide with the set of IIE mediators.

intertype compromise, one that every possible type of the principal can plausibly defend as a justifiable choice.

The intuition behind the neutral optimum in our environment can be illustrated by considering the choice between the P-Max Mediator and the Neutral Mediator. Expressing a preference for the P-Max Mediator would reveal that the principal is an L type. Conditional on this inference, an H type subordinate evaluates the proposed mediation relative to the disagreement outcome (war). Because war yields a high payoff to the H type, the subordinate is willing to immediately trigger war rather than participate in a mediator that is selected precisely because it is favored by the L type. Thus, once the principal is inferred to be L, the P-Max Mediator becomes vulnerable to blocking.

By contrast, expressing a preference for the Neutral Mediator suggests that the principal is an H type. Conditional on this inference, the relevant benchmark for the subordinate is again war, which yields a low payoff to an L type. Importantly, any higher payoff that might be achieved through strategic manipulation (i.e., misreporting) within the mediator requires the principal's continued participation and therefore cannot serve as a disagreement point. If the principal is indeed the H type, she can therefore justify selecting the Neutral Mediator by advancing the argument of the following sort: "*If you infer from my preference for the Neutral Mediator that my type is H, then we should dispense with both of these mediators and we can just directly agree on a division of the surplus that would be strictly better for you than your war payoff and just as good for me when I am the H type (e.g.,* $(240, 160)$*).*" An L type principal cannot make a symmetric argument in favor of the P-Max Mediator. Given that an H type subordinate's war payoff is already high, there is no feasible division that would make the subordinate strictly better off than war while leaving the L type principal no worse off than under the P-Max Mediator.

The burden of maintaining inscrutability falls asymmetrically on the L type, who must mimic the H type's behavior in order to avoid revealing private information. In this sense, the H type is relatively willing to reveal its type, whereas the L type has a stronger incentive to conceal it. Hence, the P-Max Mediator can be blocked, whereas the Neutral Mediator cannot. This asymmetry underlies the neutral optimum prediction: when informed principals must form an inscrutable intertype compromise, the mediator preferred by the H type emerges as the most defensible pooling choice. This reasoning leads to the following hypothesis.

Hypothesis 3 (Informed Mediator Selection: Neutral Optimum)**.** *Informed principals choose the Neutral Mediator over the P-Max Mediator.*

While Hypothesis 3 provides a sharp prediction for mediator choice, evaluating the theory of neutral optimum also requires understanding the behavioral mechanisms that underlie mediator selection. To this end, we consider two questions:

(1) what inferences subordinates draw from the principal's mediator choice; and

(2) whether subordinates decline mediation conditional on the chosen mediator.

These questions are motivated by the implicit driving force behind the theory of neutral optimum: the P-Max Mediator is not expected to be selected precisely because its selection would be declined by H type subordinates once they update their beliefs that the principal is an L type. In theory, however, subordinates should maintain their prior beliefs about the principal's type and should not decline any chosen mediator. Examining subjects' inferences and responses would allow us to assess whether the strategic logic of Myerson's theory operates in the lab.

4 Experimental Design and Procedure

We conduct two complementary experiments to test the hypotheses.

In Experiment I, subjects, depending on the treatment they belong to, play either the uninformed or the informed mediator selection game. Examining data on subjects' mediator choices allows us to test Hypotheses 1, 2(a), and 3. However, choice data alone cannot reveal the underlying sources of deviations, if any, from pooling. In particular, type-dependent mediator choices may reflect either (or both) a failure to recognize the need for inscrutability or disagreement over which mediator constitutes the appropriate intertype compromise.

Experiment II is designed to address this distinction. In this experiment, informed principals choose a mediator-selection rule that specifies not only their realized type's mediator choice but also the mediator that would be selected by their unrealized type. The objective is to examine whether informed principals choose inscrutably and, if so, to *elicit their views*—in an incentivized manner—*about how their unrealized type should behave when forming an intertype compromise*.[17] This design allows us to test Hypothesis 2(b) and to examine the behavioral foundations underlying the success or failure of the neutral optimum prediction.

The *Experimental Instructions* (url-linked) for both experiments are available online. We describe below the experimental design and procedure for each experiment. The parameterization of the model used in the experiments is provided in Section 2.3.

[17] A conventional design employing the strategy method (Selten, 1967) does not achieve this objective, as it conditions choices only on the realized type and does not require principals to explicitly consider the behavior of their unrealized type. The notion of intertype compromise instead requires a thought process in which each type of informed principal contemplates how mediator selection should be framed across types. Moreover, in a Bayesian game with type-contingent strategies, standard experimental procedures with randomly drawn types in each round are effectively equivalent to the strategy method.

4.1 Experiment I

The treatment variables are the prior probability of the H type ($q = 1/4$ or $2/5$) and the information structure at the time of mediator selection (uninformed or informed). Table 3 summarizes our 2×2 treatment design. Treatments under $q = 2/5$ are labeled "Complex" because the mediators differ in both p_H and p_M, whereas those under $q = 1/4$ are labeled "Simple" because the mediators differ only in p_M.

Table 3: Experimental Treatments in Experiment I

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

Our experiment was conducted 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 university's graduate and undergraduate population. Upon arrival at the designated Zoom meeting, subjects were instructed to turn on their videos throughout the entire course of the experiment.

We conducted four sessions for each treatment. Each subject participated in one of the 16 ($= 4 \times 4$) sessions, with session sizes ranging from 16 to 20 participants. Each session consisted of four separate parts. Parts 1–3 comprised 4 rounds each and Part 4 comprised 28 rounds, for a total of 40 rounds per session. In each round, the random matching protocol was used and subjects were independently assigned types according to q.

Part 1 corresponds to the unmediated communication game described in online Appendix D.1. Parts 2 and 3 correspond to 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 in which the two mediators appeared in Parts 2 and 3 to treat the two mediators symmetrically. Parts 1–3 served to familiarize subjects with the environment and the functioning of mediation under different mediators. As these parts are not central to our analysis, details of the experimental procedures and results are relegated to online Appendix D.1.

Part 4 implements the main treatment, in which subjects play the mediator selection game described in Section 2.2: either the uninformed mediator selection game under the Uninformed treatments or the informed mediator selection game under the Informed treatments. We describe below the detailed experimental procedure for Part 4, which generates the key observations used to test our hypotheses.

Informed Mediator Selection. This treatment was presented as Part 4 in eight experimental sessions, four under Simple-Informed and another four under Complex-Informed. At the start of each round, subjects were randomly and anonymously matched in pairs and independently assigned types by the computer according to q. After learning their own type, subjects were asked to choose which mediator of the two mediators to rely on for the round, imagining themselves as being the selector of a mediator.[18] One of the two submitted choices within each pair was then selected randomly with equal probability. The selector for the round and the mediator she has chosen were announced. The non-selector was then asked to report her inference about the selector's type based on the chosen mediator. Next, both subjects sent a confidential message from $\{H, L\}$ to the chosen mediator, except that the non-selector had an additional option to decline the mediator. If the non-selector declined, disagreement occurred immediately and payoffs were determined by the subjects' true types. Otherwise, given the reported types, the mediator prescribed agreement or walked out, and payoffs were realized. At the end of each round, subjects received feedback on both subjects' types and mediator choices, the selector and the selector's mediator choice, whether the non-selector declined, messages sent (if any), the final outcome, and their own payoff.

With regard to inference reporting, our objective was to allow non-selector subjects to update their beliefs about the selector's type based on the mediator choice. They were asked to report their posterior belief by choosing among three options: (i) more likely to be H than the prior, (ii) the same as the prior, or (iii) more likely to be L than the prior. Because belief formation is not an explicit stage of the game, this elicitation was not incentivized and is therefore treated as supplementary data. Nonetheless, it provides useful evidence on whether subjects perceive mediator choice as revealing information and helps interpret their subsequent behavior in the mediation stage.[19]

Uninformed Mediator Selection. This treatment was presented as Part 4 in eight experimental sessions, four under Simple-Uninformed and another four under Complex-Uninformed. At the start of each round, subjects were randomly and anonymously matched in pairs. Subjects were asked to choose which mediator to rely on for the round, imagining themselves as the selector; however, they did so before learning either their own type or their partner's type. Subjects knew only that each type was H or L according to q. One

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

[19] We do not anticipate subjects misreporting their beliefs, despite the absence of direct financial incentives for belief reporting. Although incentive-compatible mechanisms could have been employed after each round, Burdea and Woon (2022) highlight that the quality of reported beliefs may hinge more on the understandability of the elicitation task and the cognitive effort stimulated by incentives, rather than on formal incentive compatibility. Furthermore, Danz, Vesterlund and Wilson (2022) find that disclosing information about incentives for belief reporting can lead to a tendency to report beliefs closer to the average.

of the two submitted choices within each pair was then selected randomly with equal probability. The selector and her chosen mediator were announced, after which both subjects learned their private types, independently assigned by the computer according to q. All subsequent stages—inference, message sending (or declining), outcome and payoff realization, and feedback—followed the same procedures as in Informed Mediator Selection.

For all sessions, each subject's payoff in each round was denominated in points. One round out of 40 rounds was randomly selected by the computer to determine payment. The total payment in HKD was the points earned in the selected round, converted at a fixed rate of HK$1 per point, plus a HKD 40 show-up fee. On average, subjects earned HKD 220 ($\approx$ USD 28.20) for participating in a session that lasted approximately 1.5 hours. Payments were made electronically via the HKUST Autopay System to the bank account provided by each participant through the Student Information System.

4.2 Experiment II

Experiment II focuses on the Simple-Informed case, that is, informed mediator selection with $q = 1/4$. As in Experiment I, the experiment was conducted in English using oTree in real-time online mode via Zoom (with videos on) at HKUST. A total of 108 subjects were recruited.

Unlike Experiment I, subjects were presented with three *mediator-selection rules* instead of choosing directly between the two mediators. Each rule specified the mediator choice for both the selector's realized type and her unrealized type:

- Uncompromising rule: choose the 0-Mediator if the selector is H and the 40-Mediator if the selector is L.[20]
- P-Max Compromising rule: choose the 40-Mediator regardless of type.
- Neutral Compromising rule: choose the 0-Mediator regardless of type.

In the lab, these rules were presented without reference to their theoretical interpretations.

We conducted six sessions. Each subject participated in one session, with 18 participants per session. Each session consisted of four separate parts. Parts 1–3 comprised 5 rounds each and Part 4 comprised 25 rounds, for a total of 40 rounds per session. In each round, subjects were randomly matched in pairs; within each pair, subjects were randomly assigned, with equal probability, to the role of the selector or the non-selector and were independently assigned types according to $q = 1/4$ for that round.

[20] Another possible uncompromising option would be to choose 0-Mediator if the selector is L and 40-Mediator if the selector is H. We exclude this option because it is clearly unreasonable, as the 0-Mediator is preferred by the H type and the 40-Mediator by the L type.

Parts 1–3 were designed to familiarize subjects with the implications of the three mediator-selection rules. In each part, one of the three rules was given and fixed throughout the part. The given mediator-selection rule automatically assigned the mediator choice as a function of the selector's realized type. When the given rule was P-Max Compromising or Neutral Compromising, the resulting mediator choice was independent of type, fixed at the 40-Mediator or the 0-Mediator, respectively, making these parts equivalent to Parts 2 and 3 of Experiment I. By contrast, under the Uncompromising rule, the mediator choice depended on the selector's realized type, allowing subjects to explicitly experience type-contingent mediator selection. As they are not of central interest, details of the experimental procedures and results for Parts 1–3 are relegated to online Appendix D.2. We now describe the experimental procedure for Part 4, from which the key observations are drawn.

Informed Rule Selection. At the start of each round, subjects were randomly and anonymously matched in pairs. Within each pair, one subject was randomly chosen, with equal probability, and announced to be the selector for that round. Subjects were then independently assigned private types by the computer according to $q = 1/4$. After learning her own type, the selector was asked to choose one mediator-selection rule, among the three available options, to be used to select a mediator. The selector's mediator choice was then automatically assigned based on the chosen rule and her realized type, and the resulting mediator choice was announced. At this stage, the non-selector observed only the announced mediator and did not observe which mediator-selection rule had been chosen. The mediation process then proceeded as in Part 4 of Experiment I. At the end of each round, subjects received feedback on the selector's chosen mediator-selection rule, both subjects' realized types, the selector's mediator choice (according to the mediator-selection rule), whether the non-selector declined, messages sent (if any), the final outcome, and their own payoff.

As in Experiment I, the belief-reporting procedure for non-selectors was identical and unincentivized. We again treat it as supplementary data; yet it helps interpret our results.

Importantly, for the selector, the mediator choice specified for her *unrealized type* in each mediator-selection rule was made payoff-relevant through the following bonus-payment scheme.[21] At the end of the experiment, we randomly selected one round (from this part) in which the selector's unrealized type in the current round was in fact realized. The selector received a bonus payment of 20 points if the mediator choice specified for that unrealized type in the chosen mediator-selection rule coincided with the actual mediator choice in the randomly selected round. If no round occurred in which the unrealized type was realized, the bonus was awarded unconditionally.

[21] We conducted an auxiliary experiment without the bonus-payment scheme, which yields comparable results. We discuss them in online Appendix E.

Payment procedures were identical to those in Experiment I. On average, subjects earned HKD 221 (≈ USD 28.30), including a HKD 40 show-up fee and excluding a HKD 20 bonus payment, for participating in a session that lasted approximately 1.5 hours. Out of 59 selector-subjects who were eligible for the bonus payments, 9 received bonus payments.

5 Experimental Results: Mediator Choice

We report our main experimental findings. In Section 5.1, we compare average behavior of mediator choices made by uninformed and informed subjects across all four treatments. We use data from Experiment I aggregated over all rounds in Part 4 of all sessions within the same treatment. In Sections 5.2 and 5.3, we examine average behavior and individual behavior, respectively, regarding inscrutable intertype compromise, using data from Experiment II aggregated over all rounds in Part 4 of all sessions. We continue reporting our findings in Section 6 on inferences and responses to mediator choices, using data from both Experiments I and II. Subject behavior is consistent across the Simple and Complex environments in Experiment I, and the qualitative features of the data in both experiments remain consistent whether we use all rounds or a selected subset of rounds.

In every bar graph presented in this and next sections, we show 95% confidence intervals calculated from standard errors clustered at the session level. Tables A.1 and A.2 in Appendix A report the non-parametric test results for reference. We note that, with four independent observations (sessions) for each treatment in Experiment I, the minimum attainable one-sided p-value is 0.0625 for the Wilcoxon signed-rank test. Therefore, when reporting results for Experiment I based on this test, we use the term "predominant" to describe cases in which $p = 0.0625$. For all other non-parametric tests in both experiments, we use the standard terms "significant," "marginally significant," and "insignificant" to describe cases in which $p < 0.05$, $0.05 \leq p < 0.1$, and $0.1 \leq p$, respectively.

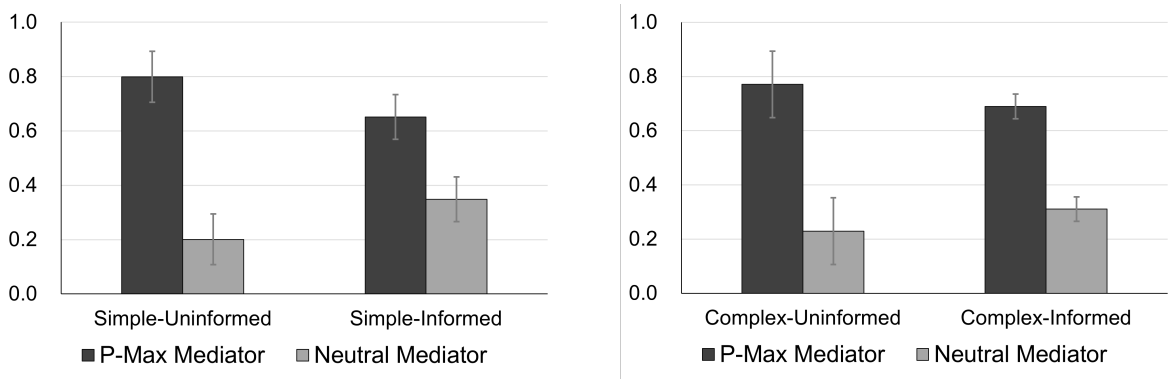
5.1 Mediator Choice

In Experiment I, subjects were unaware of their randomly assigned role as either principal or subordinate when making their mediator choice; therefore, we examine mediator choices made by all subjects rather than restricting attention to principal-subjects.

The two panels of Figure 2 report the proportions of two mediators chosen by subjects across all four treatments in Experiment I.[22] In each panel, data from the Uninformed treat-

[22]Figure B.1 in Appendix B.1 reports the proportions of two mediators chosen by principal-subjects that are subsequently played, taking into account declines by subordinate-subjects, across the four treatments. The qualitative pattern is essentially identical to that in Figure 2.

ment are shown on the left, and data from the Informed treatment on the right. Although the two proportions are complementary and thus contain the same information, we plot both bars to facilitate visual comparison across treatments.



(a) Simple (b) Complex

Figure 2: Proportion of Mediator Chosen (All Subjects)

It is immediately evident that, in the Uninformed treatments, subjects 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, 80% and 77% of subjects choose the P-Max Mediator in Simple-Uninformed and Complex-Uninformed, respectively. These observations lead to the following finding.

Finding 1 (Uninformed Mediator Selection)**.** *Consistent with Hypothesis 1, uninformed principals choose the P-Max Mediator more often than the Neutral Mediator.*

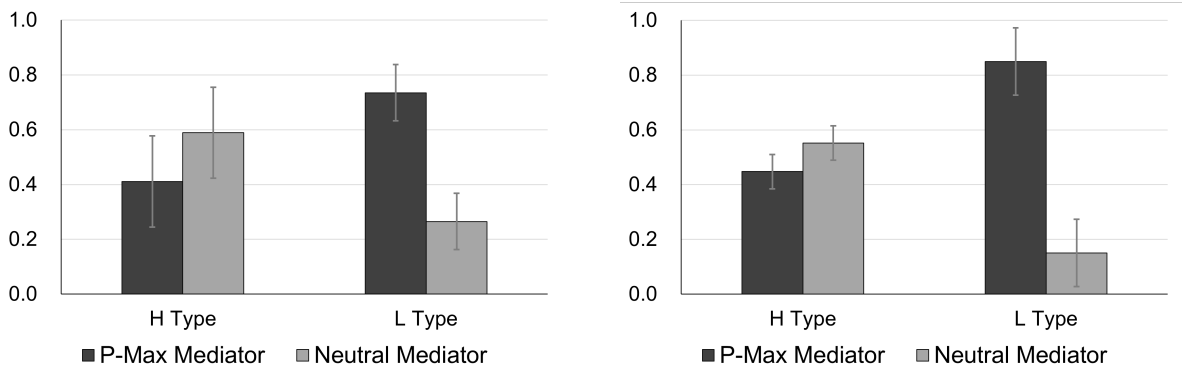
This finding suggests that, on average, subjects correctly perceive the P-Max mediator to be ex ante preferable when mediator selection occurs before learning one's type.

Turning to informed mediator selection, the Informed treatment data shown on the right side of each panel in Figure 2 indicate that subjects do not choose the Neutral Mediator over the P-Max Mediator. Instead, they choose the P-Max Mediator predominantly more often: 65% in Simple-Informed and 69% in Complex-Informed. At the same time, there is a treatment effect: Informed subjects choose the Neutral Mediator significantly more often than uninformed subjects do in both Simple and Complex treatments (35% vs. 20% and 31% vs. 23%, respectively; one-sided $p < 0.05$, Mann-Whitney tests). We summarize these observations as follows.

Finding 2 (Informed Mediator Selection)**.** *Informed principals choose the Neutral Mediator more often than do uninformed principals. Counter to Hypotheses 3, however, informed principals choose the P-Max Mediator more often than the Neutral Mediator.*

This finding indicates that average behavior in informed mediator selection deviates from the prediction of the theory of neutral optimum. We explore the underlying mechanisms behind this deviation in Section 6.

We next examine average behavior by type. Figure 3 reports the proportions of two mediators chosen by subjects of each type in the two Informed treatments. As in Figure 2, we show both bars for ease of comparison.



(a) Simple-Informed (b) Complex-Informed

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

The proportion of L type subjects choosing the P-Max Mediator is predominantly higher than that of H type subjects: 74% vs. 41% in Simple-Informed, and 85% vs. 45% in Complex-Informed. This reveals a clear type-dependent pattern in mediator choice. The pattern, however, is *asymmetric*: L type subjects choose the P-Max Mediator predominantly more often than the Neutral Mediator, whereas H type subjects choose the Neutral Mediator more often than the P-Max Mediator but the difference is statistically insignificant. These observations lead to the following finding.

Finding 3 (Informed Mediator Selection by Type)**.** *Counter to Hypotheses 2(a), the two principal types do not choose the same mediator on average. In particular, L type principals choose the P-Max Mediator more often, whereas H type principals choose the Neutral Mediator more often though not significantly so.*

We interpret this average type-dependent pattern as evidence that most subjects correctly recognize which mediator is preferred by each type given the prior, conforming with the notion of interim incentive efficiency. At the same time, H type principals are less inclined to select their type-preferred mediator.

Importantly, these observations do not, by themselves, contradict Hypothesis 2(b). Differences in average mediator choices across types need not directly imply a failure to enact

an inscrutable intertype compromise. The asymmetric type-dependent pattern in Figure 3 may reflect a subset of subjects' failing to recognize the benefit of inscrutable mediator selection, in which case their behavior would correspond to nonequilibrium play of separation. It may also arise even when subjects do understand the need for inscrutable selection but differ in their views about which mediator constitutes the appropriate intertype compromise. We examine this distinction next using data from Experiment II, which is designed to test whether and how subjects implicitly consider intertype compromise by prompting them to take the perspective of their unrealized type.

5.2 Inscrutable Intertype Compromise

In Experiment II, subjects were assigned the role of principal or subordinate at the beginning of each round; hence, mediator-selection rule choices were made only by principal-subjects.

Figure 4 reports the proportions of the three mediator-selection rules chosen by all principal-subjects in Experiment II.

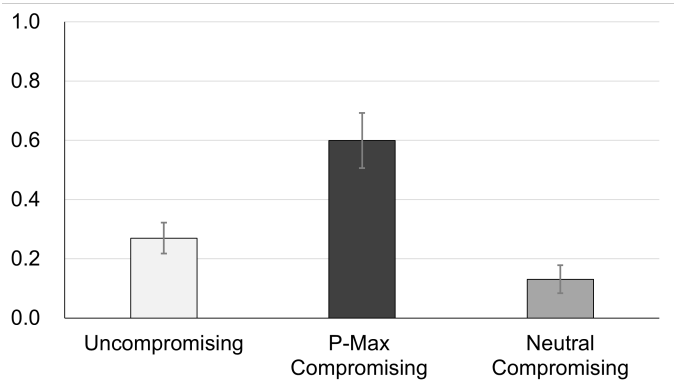


Figure 4: Proportion of Mediator-Selection Rule Chosen (All Principals)

Two observations are apparent. First, the Uncompromising rule is chosen significantly less often (27%) compared to the two Compromising rules combined (73%). Second, the P-Max Compromising rule is chosen significantly more often (60%) than the Neutral Compromising rule (13%).[23] These observations lead to the following key finding of this paper.

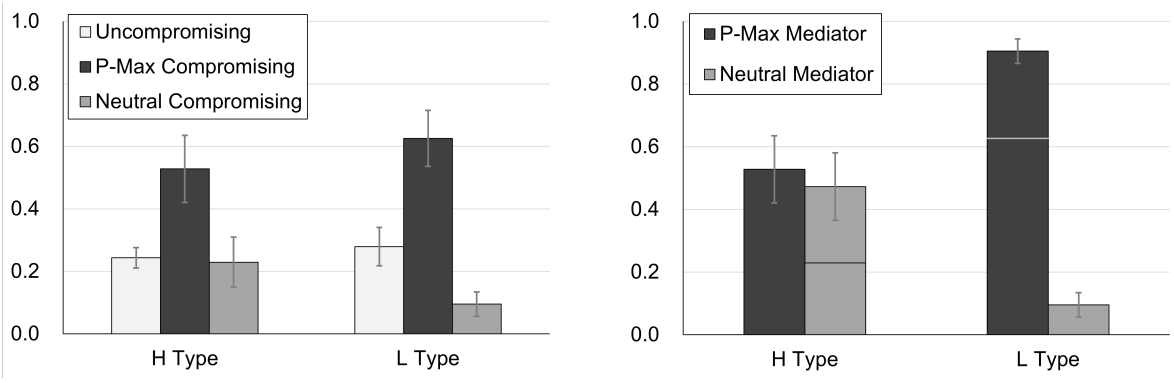
Finding 4 (Inscrutable Intertype Compromise)**.** *The majority of informed principals choose mediator-selection rules that specify the same mediator for both their realized and unrealized types, consistent with Hypothesis 2(b); and perceive the P-Max Mediator as the appropriate reasonable intertype compromise.*

[23]The Wilcoxon signed-rank tests confirm both observations (one-sided p-value= 0.0156).

The first part of this finding indicates that, on average, informed subjects understand that mediator selection should not be type-revealing, consistent with the pooling-equilibrium prediction in our setting. We interpret this as strong evidence that *subjects recognize the need for inscrutable mediator selection and engage in some form of intertype compromise between the objectives of their true type and their unrealized type.*[24] This implies that the empirically observed type-dependent pattern in Figure 3 is not a direct indication of subjects' failing to choose inscrutably.

The second part of Finding 4 reveals that, when forming such inscrutable compromises, subjects tend to resolve them in favor of the P-Max Mediator, consistent with Finding 2 in Experiment I. This pattern runs directly counter to the theory of neutral optimum, which predicts the Neutral Mediator to be the most reasonable compromise (Hypothesis 3).

Examining mediator-selection rule choices by type provides further support for this finding. Panel (a) of Figure 5 reports the corresponding proportions of choices by type.



(a) Mediator-Selection Rules Chosen (b) Resulting Mediator Choices

Figure 5: Proportion of Mediator-Selection Rule/Mediator Chosen By Type (All Principals)

Both H and L type principal-subjects choose the P-Max Compromising rule significantly more often (53% for H types and 63% for L types) than either the Uncompromising or the Neutral Compromising rule (one-sided $p < 0.03$ in either pairwise comparison for both types, Wilcoxon signed-rank tests). Among the two Compromising rules, although the discrepancy between the two is smaller for H type principals than for L type principals, the qualitative pattern is the same across types: on average, both types resolve intertype compromise toward the P-Max Mediator; this is summarized below.

[24] An alternative interpretation could be that subjects rarely choose the Uncompromising option not because they anticipate that it could signal private information, but because they do not see its strategic usefulness. We consider this interpretation unlikely for two reasons. First, subjects were exposed to all three options in earlier parts of the experiment and appeared to understand the implications of each rule. Second, the mediator specified for the unrealized type was incentivized through an explicit bonus-payment scheme, giving subjects clear incentives to consider both components of the mediator-selection rule.

Finding 5 (Inscrutable Intertype Compromise by Type)**.** *Both H and L type principals perceive the P-Max Mediator as the appropriate reasonable intertype compromise.*

Taken together, Findings 4 and 5 reinforce the conclusion that subjects, on average, form an implicit intertype compromise in making inscrutable selection.

Why, then, does the average pattern of mediator choices observed in Experiment I appear type-dependent (Figure 3)? The mediator-selection rule choices by type shown in Figure 5(a) clarify the source of this pattern. Once the chosen rules are mapped into the mediators they implement, it becomes clear how type-dependent mediator choices can arise even when subjects choose mediator-selection rules in an ostensibly inscrutable manner. Panel (b) of Figure 5 shows the proportions of the resulting mediator choices by type implied by the chosen rules and thus observed by subordinate-subjects. The resulting pattern closely mirrors that observed in Figure 3 and is consistent with Finding 3 in Experiment I: L type principals choose the P-Max Mediator more often than the Neutral Mediator, while H type principals' choices between the P-Max and Neutral Mediators are not statistically distinguishable.

These observations suggest that the observed type-dependent mediator choices arise from two sources. First, a similar fraction of principals of both types fail to enact an inscrutable intertype compromise, effectively adopting a type-revealing separating strategy (24% of H type principals and 28% of L type principals choose the Uncompromising rule). This behavior contributes symmetrically, on average, to the observed divergence across types by increasing the proportion of principals choosing the Neutral Mediator among H types and the P-Max Mediator among L types (depicted by the upper portions above the horizontal lines in Figure 5(b)). Thus, it can be interpreted as a common behavioral deviation from the theoretical prediction, which by itself does not fully account for the observed asymmetry.

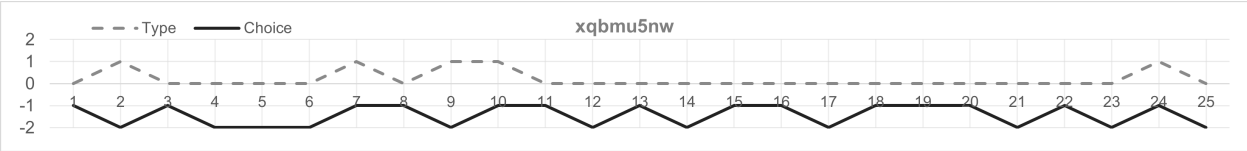
More importantly, the second source of type-dependent mediator choices concerns disagreement among principals who do appear to enact an inscrutable intertype compromise but disagree about how that compromise should be resolved. This disagreement is driven largely by heterogeneity within the H type: 23% of H type principals favor pooling on the Neutral Mediator, whereas a majority of both H types (53%) and L types (63%) favor pooling on the P-Max Mediator. Such disagreement among a subset of H type principals relative to the rest of the population, rather than a failure to recognize the need for inscrutable mediator selection itself, contributes to the asymmetric type-dependent mediator choices.

5.3 Classifications of Individual Behavior

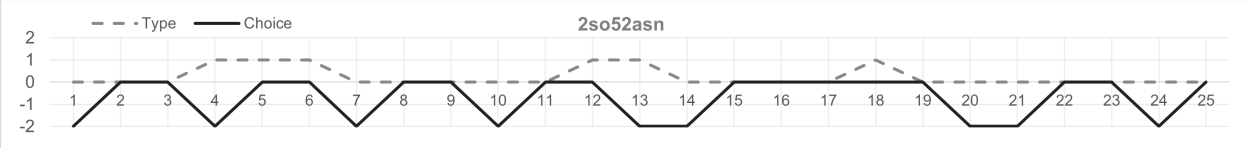
We take a closer look at mediator-selection rule choices by individuals in Experiment II. Figure 6 presents five representative individual cases, illustrating each subject's decisions as a

Principal over 25 rounds.[25] In each panel (with the subject ID displayed in the top), the horizontal axis represents rounds. On the vertical axis, dotted lines indicate type realizations (1=H, 0=L), while solid lines denote choices (1=Neutral Compromising, 0=P-Max Compromising, -1=Uncompromising, and -2=no deliberate choice as acting as a subordinate).

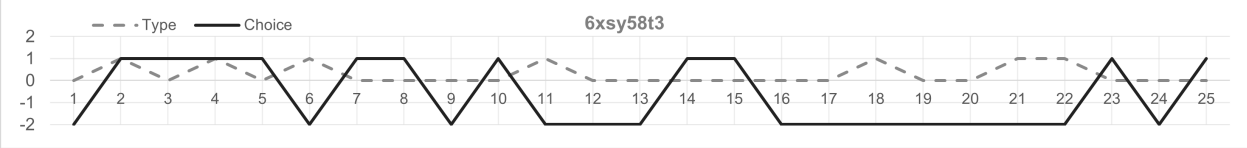
(a) Uncompromising



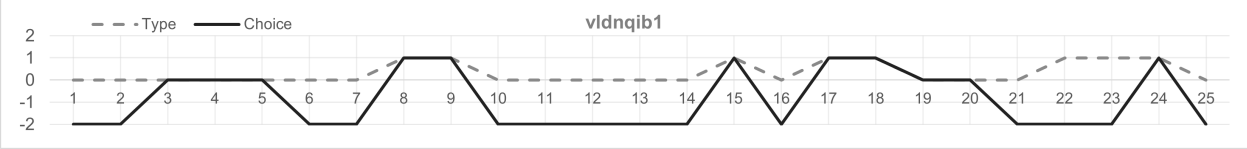
(b) P-Max Compromising



(c) Neutral Compromising



(d) Type-divergent Compromising



(e) Random

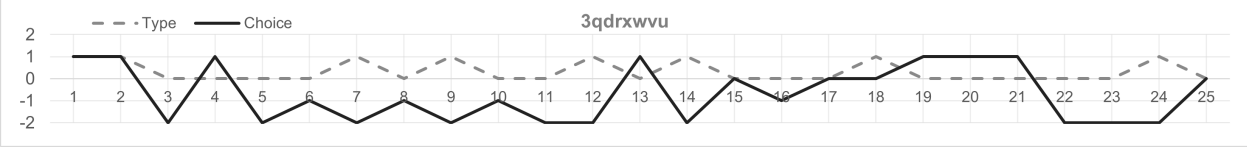


Figure 6: Five Representative Cases

Each of Panels (a), (b), and (c) shows a subject who consistently selects Uncompromising, P-Max Compromising, and Neutral Compromising, respectively, regardless of type across principal-rounds. Panel (d) displays perfectly type-divergent choices of either P-Max or Neutral Compromising. Panel (e) represents behavior that is neither consistent nor type-contingent.

[25] We have 108 such figures in total, available upon request.

We consolidate all 108 individual observations using two proxy measures for their rule choices: *inscrutability* and *type-polarization*. Inscrutability measure is quantified as the fraction of rounds in which an individual selects either the P-Max or Neutral Compromising rule. Type-polarization measure is calculated as the total variation distance between the conditional choice distribution of the two types over the three mediator-selection rules.

Classification of individuals into k clusters is performed using these two dimensions via the k-means clustering method (Macqueen, 1967). We chose $k = 5$ to obtain classifications with meaningful and interpretable behavioral patterns, as shown in Figure 7.

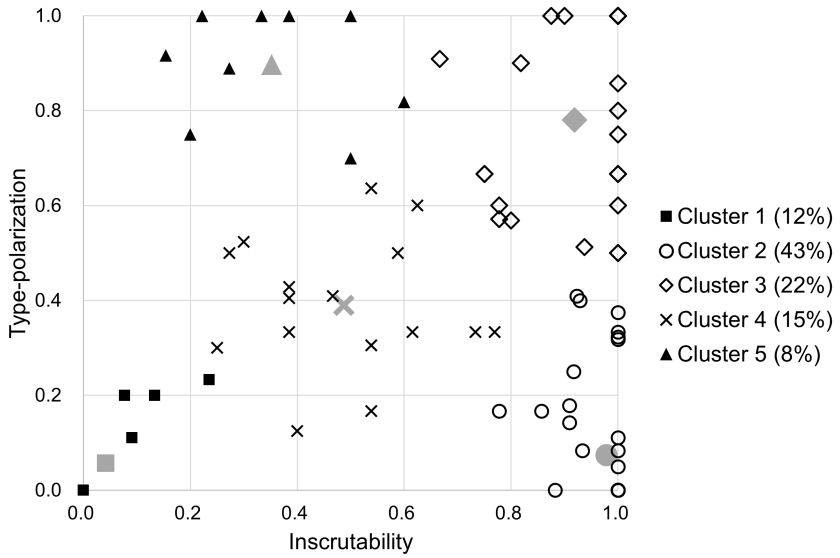


Figure 7: k-means Clustering with Five Clusters

Note: The horizontal axis measures the frequencies of inscrutable selections, represented by choices of the P-Max and Neutral Compromising rules. The vertical axis measures the distributional divergence in choice frequencies across the three mediator-selection rules between the two types. Due to multiple subjects sharing the same pair of measures, certain markers represent more than one subject, notably at $(0, 0)$, $(1, 0)$, and $(1, 1)$. Cluster centroids are indicated by the corresponding marker shapes shaded with gray.

Cluster 1 (■, 12%) consists of 13 individuals with low inscrutability and low type-polarization. These individuals mostly select the Uncompromising rule across all principal-rounds, abstaining from inscrutable selection. Among them, 9 individuals (concentrated at $(0, 0)$) always select the Uncompromising rule regardless of type, as illustrated by the representative individual in panel (a) of Figure 6.

Cluster 2 (○, 43%) consists of 46 individuals with high inscrutability and low type-polarization. These individuals dominantly select one of the two compromising rules (P-Max or Neutral) across all principal-rounds, engaging in consistent inscrutable selection and converging on a stable, type-independent intertype compromise. Among them, 29 individuals always select the P-Max Compromising rule regardless of type, while one individual always

selects the Neutral Compromising rule regardless of type (all 30 concentrated at $(1, 0)$); representative individuals are shown in panels (b) and (c) of Figure 6.

Cluster 3 ($\diamond$, 22%) consists of 24 individuals with high inscrutability and high type-polarization. These individuals dominantly select one of the two compromising rules (P-Max or Neutral) across all principal-rounds but exhibit between-type divergence, engaging in inscrutable selection without converging on a common intertype compromise between the two types. Among them, 6 individuals (concentrated at $(1, 1)$) always select a compromising rule contingent on type, generating complete divergence between types. A representative individual, shown in panel (d) of Figure 6, always chooses the P-Max Compromising rule when the realized type is L and the Neutral Compromising rule when the realized type is H.

Cluster 4 ($\times$, 15%) consists of 16 individuals with moderate inscrutability and moderate type-polarization. These individuals largely fail to internalize the need for inscrutable selection but display limited type asymmetry. When they choose compromising options, their choices often share partial common support across types, although this cluster also includes individuals with largely random choice patterns (e.g., a representative individual shown in panel (e) of Figure 6). Cluster 5 ($\blacktriangle$, 8%) consists of 9 individuals with low-to-moderate inscrutability and high type-polarization. These individuals largely fail to recognize the need for inscrutable selection, similar to Cluster 4, and exhibit pronounced type polarization.

These clustering results reinforce the main findings from the aggregate analysis. A large majority of principals (Clusters 2 and 3; 70 individuals, 65%) appear to engage in inscrutable intertype compromise. In particular, 51 individuals (47% of the sample) exhibit "Inscrutability=1," always making inscrutable selections. Within this group, 37 individuals (belonging to Cluster 2) converge on a stable, type-independent intertype compromise, while 14 individuals (belonging to Cluster 3) display type-contingent intertype compromises. By contrast, a subset of principals (Clusters 4 and 5; 25 individuals, 23%) shows limited engagement in inscrutable compromise, and a smaller group (Cluster 1; 13 individuals, 12%) shows no engagement in inscrutable compromise, exhibiting non-equilibrium behavior.

6 Why the Theory of Neutral Optimum Fails?

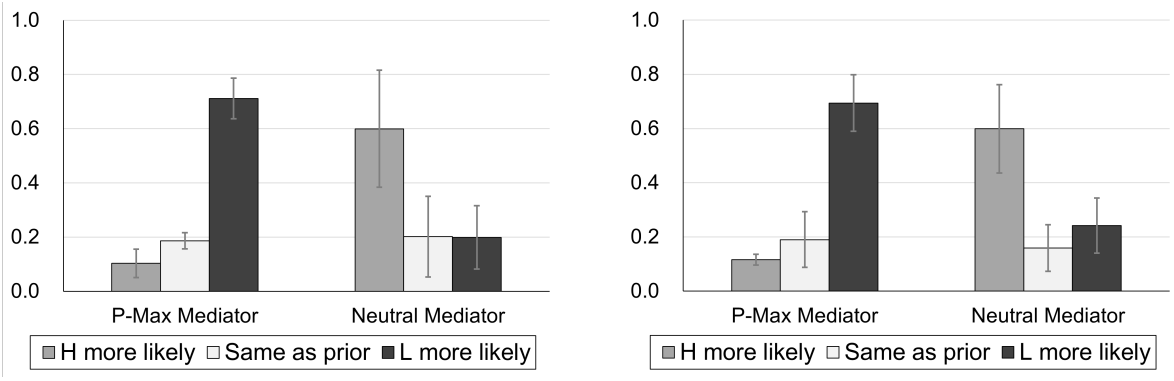
Findings 2 and 4 clearly indicate a failure of the theory of neutral optimum. We now turn to the underlying behavioral mechanisms by examining subordinates' inferences and participation strategies after the principal's chosen mediator is announced in Part 4 of Experiment I (Informed treatments) and Experiment II.[26]

[26] We focus on subordinates' responses to mediator choice, as these may in turn directly influence principals' mediator-selection decisions. For completeness, we report in Appendix B.2 data on principals' partici-

6.1 Subordinate's Inference

After the principal's chosen mediator was announced, subordinates were asked to select one of three inferences about the principal's type: (i) More likely to be H (than under the prior), (ii) Same as the prior, and (iii) More likely to be L (than under the prior).

Figure 8 reports the average frequencies of three inferences by subordinates conditional on either the P-Max Mediator or the Neutral Mediator chosen by the principal in the Informed treatments of Experiment I. In the figure, the inference categories are abbreviated as 'H more likely,' 'Same as prior,' and 'L more likely.' Subordinates infer that the principal is more likely to be the L type after observing the principal's choice of the P-Max Mediator with predominantly high frequency (71% in Simple; 69% in Complex); and that the principal is more likely to be the H type after observing the principal's choice of the Neutral Mediator with predominantly high frequency (60% in both Simple and Complex).[27]



(a) Simple-Informed (b) Complex-Informed

Figure 8: Subordinate's Inference Conditional on Mediator Choice (Experiment I: Informed)

Figure 9 reports the average frequencies of three inferences by subordinates conditional on either the P-Max Mediator or the Neutral Mediator chosen by the principal (according to her chosen mediator-selection rule) in Experiment II. We observe a pattern similar to that in Figure 8: the frequency of 'L more likely' inference following the P-Max Mediator is 63%, and the frequency of 'H more likely' inference following the Neutral Mediator is 52%, with both frequencies significantly higher than the frequencies of the other two inferences.

pation strategies that are not discussed here.

[27]The inference data exhibit a stark contrast between the Informed and Uninformed treatments in Experiment I (cf. Figure B.2 in Appendix B.1). In the Uninformed treatments, inferences do not vary asymmetrically with the chosen mediator. For each mediator, the frequencies of the responses 'same as prior' and 'L more likely' are comparable (42-43% and 48%, respectively, given P-Max; 35-37% and 29-30% given Neutral, across the Simple and Complex treatments). This pattern suggests that subordinates did not extract additional information from the mediator choice when principals were uninformed.

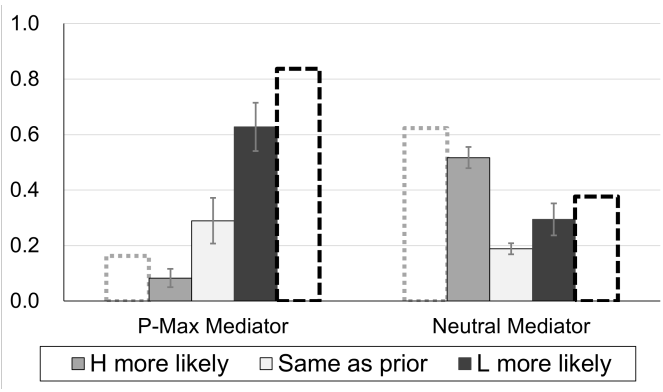


Figure 9: Subordinate's Inference Conditional on Mediator Choice (Experiment II)

Note: The bars with gray dotted borders and black dashed borders show the empirical posterior probabilities that the principal is H type and L type, respectively, conditional on each mediator choice.

The inference patterns observed in both experiments lead to the following finding, addressing the first question posed after Hypothesis 3 in Section 3.3: what inferences subordinates draw from the principal's mediator choice.

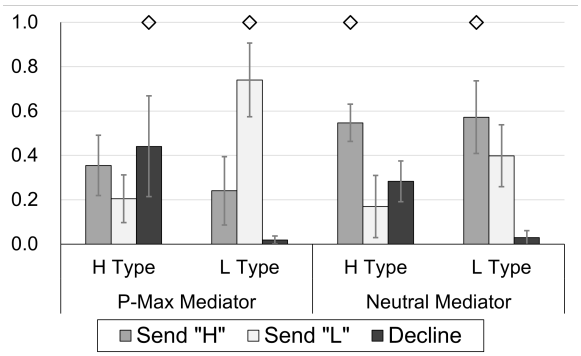
Finding 6 (Inference)**.** *Subordinates update their beliefs in response to the informed principal's mediator choice, inferring that the principal is more likely to be the L type following the P-Max Mediator and more likely to be the H type following the Neutral Mediator.*

While belief reporting was not incentivized and should therefore be interpreted as supplementary evidence only, this finding suggests that subjects systematically perceive the informed principal's mediator choice as revealing information, even when principals appear to make inscrutable selections. In theory, subordinates should maintain their prior beliefs about the principal's type. Figure 9 also depicts the empirically implied Bayesian posteriors of the principal's type following each mediator choice, calculated from the proportions of each principal type conditional on the mediator chosen. These observations indicate that principals' mediator choices affect subordinates' beliefs, with reported beliefs changing in the "correct" direction.

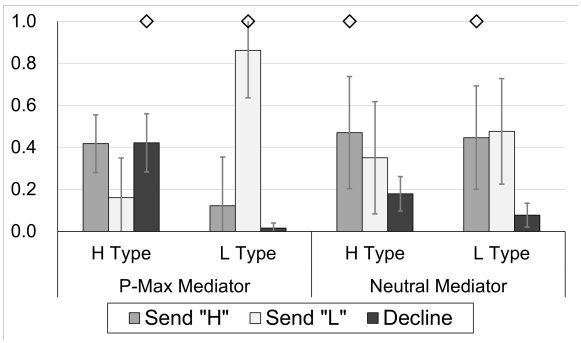
6.2 Subordinate's Participation Strategy

After the mediator announcement and belief reporting, subordinates either sent a message ("H" or "L"), in which case the mediator's recommendation was implemented based on the messages sent by both players, or declined the mediator, in which case disagreement resulted.

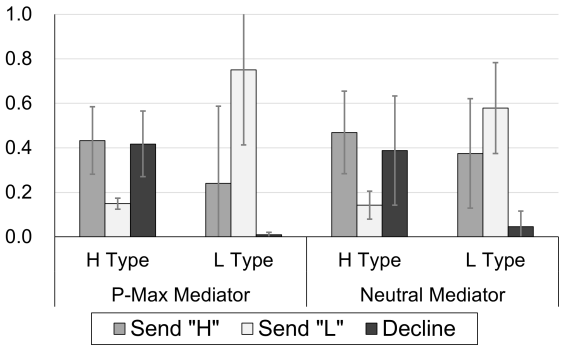
Figure 10 reports the average frequencies of subordinate strategies by type, conditional on whether the principal selected the P-Max Mediator or the Neutral Mediator, across all four treatments of Experiment I.



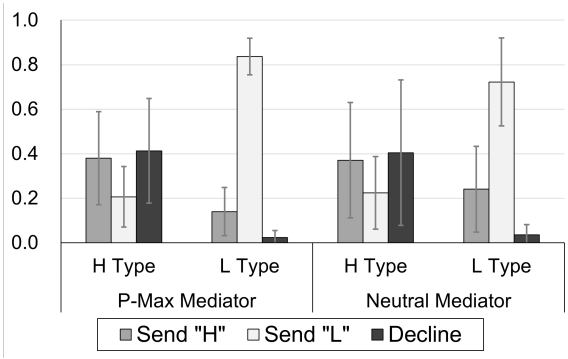
(a) Simple-Informed



(b) Complex-Informed



(c) Simple-Uninformed



(d) Complex-Uninformed

Figure 10: Subordinate's Strategy in Chosen Mediator by Type (Experiment I)

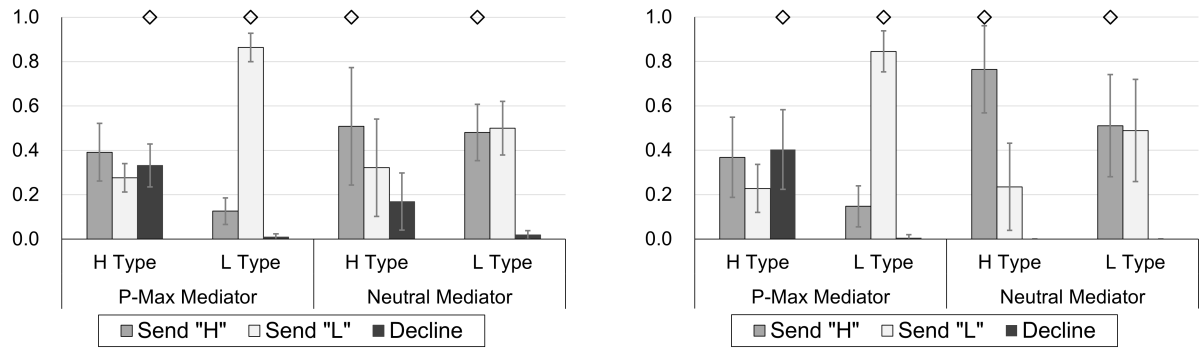
Note: The diamond markers in panels (a) and (b) indicate the subordinate's best responses for each type given updated posterior beliefs about the principal's type, specifically, $q' > q$ following the choice of the P-Max Mediator and $q' < q$ following the choice of the Neutral Mediator, which are consistent with the predominant beliefs reported on average. These best responses are computed against the principal sending truthful messages, and are formally characterized given any arbitrary posterior beliefs in Appendix B.4.

Figure 11 reports the corresponding frequencies for Experiment II, which focuses on the Simple-Informed environment. While average behavior was relatively stable across rounds in Experiment I, it exhibited noticeable variation in Experiment II; for this reason, we report data aggregated over all rounds in panel (a) and over the final five rounds in panel (b).

We abstract from scrutinizing departures of subject behavior from the theoretical predictions of pooling equilibria and optimal mediation, as our objective is not to analyze sincerity in mediation or its fragility, as in CFP. Instead, we focus on behavioral patterns that shed light on why the neutral optimum fails to emerge in the lab.

Figures 10 and 11(a) reveal several regularities in average subordinate behavior across environments. We highlight particularly salient patterns.

Comparing panels (a) and (c), and panels (b) and (d) in Figure 10, we observe a treatment effect for L type subordinates when the principal's choice is the Neutral Mediator: the



(a) All Rounds

(b) Last 5 Rounds

Figure 11: Subordinate's Strategy in Chosen Mediator by Type (Experiment II)

Note: The diamond markers have the same interpretation as in Figure 10.

frequency of "L" message by L type subordinates decreases from 58% in Simple-Uninformed to 40% in Simple-Informed, and from 72% in Complex-Uninformed to 48% in Complex-Informed. That is, L type subordinates are more truthful when an uninformed principal chooses the Neutral Mediator than when an informed principal does, although the effect is marginally significant (one-sided $p = 0.0571$ for both Simple and Complex, Mann-Whitney tests). Further, as shown in panels (a) and (b) of Figure 10 and in both panels of Figure 11, L type subordinates send truthful messages substantially more often to the P-Max Mediator than to the Neutral Mediator when chosen by the informed principal.[28] Moreover, L type subordinates hardly ever decline any mediator in any environment. Taken together, these patterns suggest that L type subjects, at least partially, update their posteriors and condition their strategies on the principal's mediator choice.

Distinct patterns emerge for H type subordinates. As shown in panels (a) and (b) of Figure 10 and in both panels of Figure 11, the frequencies with which H type subordinates decline the P-Max and Neutral Mediators are, respectively, 44% vs. 28% in Simple-Informed and 42% vs. 18% in Complex-Informed in Experiment I; and 33% vs. 17% across all rounds and 40% vs. 0% in the last five rounds in Experiment II. The Wilcoxon signed-rank tests (one-sided) confirm these differences are predominant in Experiment I and significant in Experiment II. That is, H type subordinates decline the P-Max Mediator more often than the Neutral Mediator when chosen by the informed principal. However, average decline rates are always below 45%. More strikingly, even in the Uninformed treatments of Experiment I (panels (c) and (d) of Figure 10), H type subordinates decline both mediators with similar

[28]The frequencies of "L" message to the P-Max and Neutral Mediators are, respectively, 74% vs. 40% in Simple-Informed (one-sided $p = 0.0625$) and 86% vs. 48% in Complex-Informed (one-sided $p = 0.0625$) in Experiment I; 86% vs. 50% across all rounds (one-sided $p = 0.0156$) and 85% vs. 49% in the last five rounds (one-sided $p = 0.03125$) in Experiment II. All p-values are from Wilcoxon signed-rank tests.

frequencies: 42% vs. 39% in Simple-Uninformed and 41% vs. 40% in Complex-Uninformed for P-Max and Neutral, respectively. The differences in these decline frequencies between the Uninformed and Informed treatments are statistically insignificant. Unlike L type subordinates, the H type behavior does not appear to be largely driven by conditioning on the principal's mediator choice.

Among the observations, we highlight the following finding, which addresses the second question posed after Hypothesis 3 in Section 3.3: whether subordinates decline mediation conditional on the chosen mediator.

Finding 7 (Decline Strategy)**.** *L types subordinates rarely decline any mediator, whether chosen by uninformed or informed principals. H type subordinates decline the P-Max Mediator more often than the Neutral Mediator when principals are informed, but at rates below a majority on average; and decline both mediators at similar sub-majority rates when principals are uninformed.*

Average subordinate behavior is qualitatively similar across the Informed and Uninformed treatments of Experiment I, as well as across Experiments I and II, and exhibits little systematic change. These patterns indicate that although subordinates' reported beliefs respond to mediator choice (see Figures 8 and 9), the behavioral implications of these belief updates are limited, suggesting that any change in beliefs is small.

In theory, absent belief updating, subordinates should never decline mediation. In all panels of Figures 10 and 11, we observe a lack of optimal behavior on average under the assumption that subjects hold prior beliefs. Even when taking into account that subjects may infer the principal to be an H type following the choice of the Neutral Mediator and an L type following the choice of the P-Max Mediator, subjects lack optimal behavior. Given such changes in posterior beliefs from the priors, L type subordinates should optimally misreport in the P-Max Mediator, and H type subordinates should optimally decline the P-Max Mediator and never decline the Neutral Mediator, as indicated by the diamond markers in Figures 10 and 11. The observed behavior departs from both benchmarks. Instead, L type subordinates exhibit insincerity when their less preferred mediator is chosen, regardless of whether the principal is informed or uninformed; H type subordinates display a general propensity to decline mediation regardless of the chosen mediator or the informational environment.

6.3 Discussion of Failure of Neutral Optimum Theory

The behavioral patterns observed in our experiments suggest that subjects have some understanding of the underlying structure of the game, insofar as they tend to choose their

type-preferred mediator and make corresponding inferences about the principal's type; however, these (unincentivized) inferences appear to have little effect on subjects' subsequent plays in mediation, and hence on mediator selection itself. In particular, subjects place less weight on fine-tuning their strategies in response to mediator choices or strategic inferences: L types only partially lie in the Neutral Mediator, and H types do not decline the P-Max Mediator as much.

Importantly, evidence from Experiment II indicates that H type subordinate behavior evolves over time. While H types' responses to the P-Max Mediator remain relatively consistent across rounds, their behavior under the Neutral Mediator changes markedly (cf. H type behavior under Neutral Mediator in panels (a) and (b) of Figure 11). In the final five rounds, H type subordinates become substantially more truthful and never decline the mediator. We interpret this convergence in later rounds, together with H types' tendency to decline mediation much less often than predicted under full inference, as the key source of the failure of the neutral optimum in the lab.

The theory of neutral optimum predicts that the Neutral Mediator constitutes the most reasonable intertype compromise. This prediction crucially relies on an asymmetric strategic structure. As discussed in Section 3.3, an H type principal is in a stronger position with respect to type revelation relative to the disagreement point. Even if the principal were revealed to be the H type, an L type subordinate's misreporting in mediation would not constitute a credible threat, because the principal could offer outcomes that dominate disagreement, thereby deterring unfavorable inferences by the subordinate. By contrast, if the principal were revealed to be the L type, an H type subordinate could credibly threaten to enforce disagreement, as the L type principal cannot offer any outcome strictly better than disagreement. For the Neutral Mediator to arise as an intertype compromise, principals must internalize this asymmetry when forming their compromise prior to mediator selection.

Our experimental evidence suggests that subjects neither made compromises in this manner nor learned to do so over time. In particular, H type subordinate behavior fails to generate the strategic pressure necessary for such a compromise to arise. While both types behave relatively optimally by being truthful when their preferred mediator is chosen, L type subordinates on average find ways to gain by misreporting in the Neutral Mediator. At the same time, H type subordinates exhibit a tendency to decline mediation that is too weak to render the choice of P-Max Mediator sufficiently costly for L type principals. As a result, L type principals do not experience the adverse consequences associated with choosing the P-Max Mediator and thus face little incentive to adopt an implicit compromise by selecting the Neutral Mediator; instead, they choose to misreport when the Neutral Mediator is chosen. Likewise, H type principals do not incur a sufficiently large loss from choosing the

P-Max Mediator relative to their preferred Neutral Mediator, given that L types subordinates tend to misreport in the Neutral Mediator, which is harmful for H types. As such, subjects converge on an intertype compromise tilted toward the P-Max Mediator, selecting it inscrutably, thus the theory of neutral optimum fails in the lab.

7 Concluding Remarks

We experimentally study the informed principal's mechanism selection problem. To our knowledge, this is the first attempt of a lab experiment on informed principal problems. Consistent with theory, uninformed subjects predominantly choose the ex ante peace-maximizing mediator. However, informed subjects do not choose the neutral mediator over the peace-maximizing mediator, contrary to the prediction of the theory of neutral optimum. Our evidence indicates that informed subjects recognize the need for inscrutable selection, making some inscrutable intertype compromise. When incentivized to consider the perspective of their unrealized type, both types of subjects perceive the ex ante peace-maximizing mediator as the focal compromise. As a result, the neutral optimum does not occur, nor is there evidence of learning toward it over repeated interactions. Although subjects update their beliefs in response to mediator choice, these belief updates have little effect on their subsequent strategies. In particular, high type subordinates do not decline the peace-maximizing mediator frequently enough to impose meaningful costs on low type principals. Because this blocking logic fails to operate, low types face limited pressure to concede to the high type preferred neutral mediator. As a result, the strategic asymmetry underlying the neutral optimum does not materialize in the lab.

Myerson (1983) considers the informed principal's mechanism selection as part of a non-cooperative game, which in principle can be implemented in the lab. However, the strategic logic underlying theoretical concepts, such as intertype compromise or the neutral optimum, may be difficult for subjects to fully internalize. To assess the role of incentives in facilitating this reasoning, we conducted an auxiliary experiment using the same design as Experiment II but without making the unrealized-type's choice payoff-relevant (see online Appendix E). Even in this setting, informed subjects continued to select pooling rules, and both types made an inscrutable intertype compromise. However, while a majority of low types still resolved the compromise toward the peace-maximizing mediator, there was greater disagreement among high type principals: a larger fraction favored pooling on the neutral mediator relative to the main experiment. This pattern suggests that although subjects broadly grasp the logic of inscrutable selection, the formation of an intertype compromise depends sensitively on explicit incentives. In the absence of such incentives, subjects appeared less

inclined to carry out the more sophisticated internal deliberation required to form a stable compromise, and sought divergent compromises. At the same time, subject behavior in our experiments remains far from erratic. Subjects correctly identify which mediator benefits which type, update their beliefs in response to mediator choice in the appropriate direction, and respond strategically in the mediation stage. Thus, even in a complex environment, behavior broadly aligns with the strategic structure emphasized by the informed principal problem, although not in a way that fully sustains the neutral optimum.

We conclude by discussing potential directions for future research. A fundamental challenge is that it is difficult to design a laboratory environment in which mediator selection is completely insulated from information revelation and the strategic responses it induces. In our experiments, subjects formed inferences and strategically chose their actions, yet they did not appear to fully incorporate these considerations into their initial mediator choices. One possibility is that subjects perceive their inferences as imperfect, or that belief changes are too small to justify substantial adjustments in mediator selection. Providing subjects with information during their decision stages about others' type-contingent choices may strengthen the link between inference and strategic behavior. Another possible explanation concerns the complexity of the game itself. The informed principal's mechanism selection problem combines signaling considerations with the intrinsic complexity of mechanism design under incomplete information. An important question is how this complexity shapes signaling through mediator choice. Future work could consider simpler and more controlled designs. For example, an environment that systematically enforces truthful implementation of the chosen mediator. In such a design, after the principal selects a mediator, each subject would simply choose between "in" and "out," and if both chooses "in," payoffs would be realized according to the truthful implementation of the mediator's plan, without additional inferences or reporting stages. Alternatively, the informed principal problem could be examined in a simpler sender-receiver framework only with obedience constraints.

Appendix A: Non-parametric Test Results

Table A.1: **Non-parametric Tests for Experiment I (Part 4 in the Simple and Complex Treatments)**

Reference	Test	Null Hypothesis	p-values	
			Simple	Complex
Fig. 2 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
Fig. 2 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
	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. B.1	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 Finding 3	WSR	3.1 (Informed, all subjects) The rate of choosing P-Max is the same across the two types.	0.125	0.125
	WSR	3.2 (Informed, all subjects) The rate of choosing P-Max = the rate of choosing Neutral, by H type.	0.25	0.25
	WSR	3.3 (Informed, all subjects) The rate of choosing P-Max = the rate of choosing Neutral, by L type.	0.125	0.125
Fig. 8	WSR	4.1 (Informed) Given P-Max chosen, the rate of 'L' inference = the rate of 'Same' inference.	0.125	0.125
	WSR	4.2 (Informed) Given P-Max chosen, the rate of 'L' inference = the rate of 'H' inference.	0.125	0.125
	WSR	4.3 (Informed) Given Neutral chosen, the rate of 'H' inference = the rate of 'Same' inference.	0.125	0.125
	WSR	4.4 (Informed) Given Neutral chosen, the rate of 'H' inference = the rate of 'L' inference.	0.125	0.125
	WSR	5.1 (Informed) The rate of 'L' inference given P-Max = the rate of 'L' inference given Neutral.	0.125	0.125
Fig. B.2	WSR	5.2 (Uninformed) The rate of 'L' inference given P-Max = the rate of 'L' inference given Neutral.	0.125	0.125
	WSR	5.3 (Uninformed) The rate of 'Same' inference given P-Max = the rate of 'Same' inference given Neutral.	0.125	0.625

(continued on the next page)

Table A.1: **(Continued)**

Reference	Test	Null Hypothesis	p-values	
			Simple	Complex
Fig.8 & Fig. B.2	MW	5.4 The rate of 'Same' inference given P-Max is the same across Uninformed and Informed.	0.0286	0.0571
	MW	5.5 The rate of 'Same' inference given Neutral is the same across Uninformed and Informed.	0.0286	0.0286
Fig. 10	WSR	6.1 (Informed) The rate of rejecting P-Max = the rate of rejecting Neutral, by H type.	0.125	0.125
	WSR	6.2 (Informed) The rate of rejecting P-Max = the rate of rejecting Neutral, by L type.	0.875	0.125
	WSR	6.3 (Uninformed) The rate of rejecting P-Max = the rate of rejecting Neutral, by H type.	0.875	1
	WSR	6.4 (Uninformed) The rate of rejecting P-Max = the rate of rejecting Neutral, by L type.	0.25	0.625
	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	0.3429
	WSR	8.1 (Informed) The rate of "H" message in P-Max = that in Neutral, by H type.	0.125	0.375
	WSR	8.2 (Informed) The rate of "L" message in P-Max = that in Neutral, by L type.	0.125	0.125
	WSR	8.3 (Uninformed) The rate of "H" message in P-Max = that in Neutral, by H type.	0.625	1
	WSR	8.4 (Uninformed) The rate of "L" message in P-Max = that in Neutral, by L type.	0.25	0.125
	MW	9.1 The rate of "H" message in P-Max by H type is the same across Uninformed and Informed.	0.6857	0.6857
	MW	9.2 The rate of "H" message in Neutral by H type is the same across Uninformed and Informed.	0.3429	0.6857
	MW	9.3 The rate of "L" message in P-Max by L type is the same across Uninformed and Informed.	0.6857	0.3429
	MW	9.4 The rate of "L" message in Neutral by L type is the same across Uninformed and Informed.	0.1143	0.1143

■ WSR and MW refer to the Wilcoxon signed-rank test and Mann-Whitney test, respectively. All null hypotheses are two-sided.
■ With four independent sessions for each treatment, the minimum attainable p-value is 0.125 (two-sided) for the Wilcoxon signed-rank test.
■ P-Max and Neutral refer to the P-Max and Neutral Mediators, respectively.
■ In parenthesis, we indicate which treatment (Uninformed or Informed) data we consider and/or whether we test for all subjects or principal-subjects. For 4.1–9.4, we test for only subordinate-subjects.
■ In 4.1–5.5, inference labels are abbreviated ('L'='More likely to be L'; 'Same'='Same as prior'; 'H'='More likely to be H').

Table A.2: **Non-parametric Tests for Experiment II (Part 4)**

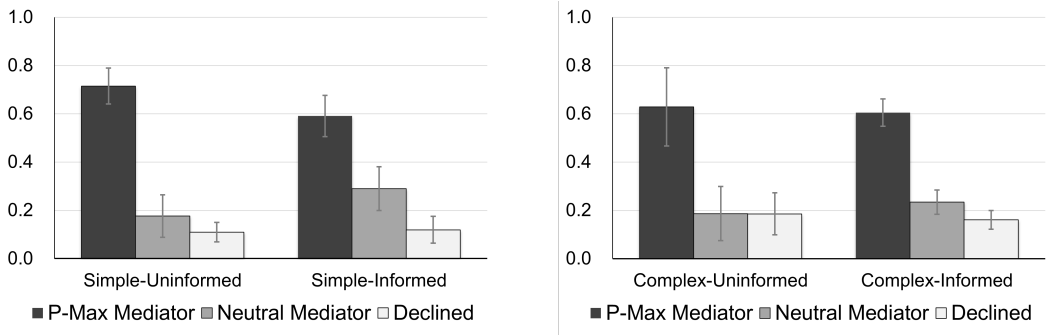
Reference	Test	Null Hypothesis	p-values
Fig. 4 Finding 4	WSR	1.1 The rate of choosing P-Max Comp. = the rate of choosing Uncompromising.	0.03125
	WSR	1.2 The rate of choosing P-Max Comp. = the rate of choosing Neutral Comp.	0.03125
	WSR	1.3 The rate of choosing Compromising (both P-Max and Neutral) = the rate of Uncompromising.	0.03125
Fig. 5	WSR	2.1 For H type, the rate of choosing P-Max Comp. = the rate of choosing Uncompromising.	0.03125
	WSR	2.2 For H type, the rate of choosing Neutral Comp. = the rate of choosing Uncompromising.	0.6875
	WSR	2.3 For H type, the rate of choosing P-Max Comp. = the rate of choosing Neutral Comp.	0.05906
	WSR	2.4 For L type, the rate of choosing P-Max Comp. = the rate of choosing Uncompromising.	0.03125
	WSR	2.5 For L type, the rate of choosing Neutral Comp. = the rate of choosing Uncompromising.	0.03125
	WSR	2.6 For L type, the rate of choosing P-Max Comp. = the rate of choosing Neutral Comp.	0.03125
Fig. 9	WSR	3.1 Given P-Max announced, the rate of 'L' inference = the rate of 'Same' inference.	0.03125
	WSR	3.2 Given P-Max announced, the rate of 'L' inference = the rate of 'H' inference.	0.03125
	WSR	3.3 Given Neutral announced, the rate of 'H' inference = the rate of 'Same' inference.	0.03125
	WSR	3.4 Given Neutral announced, the rate of 'H' inference = the rate of 'L' inference.	0.03125
Fig. 11	WSR	4.1 (All rounds) The rate of rejecting P-Max = the rate of rejecting Neutral, by H type.	0.03552
	WSR	4.2 (All rounds) The rate of rejecting P-Max = the rate of rejecting Neutral, by L type.	1
	WSR	5.1 (All rounds) The rate of "H" message in P-Max = that in Neutral, by H type	0.4375
	WSR	5.2 (Last five rounds) The rate of "H" message in P-Max = that in Neutral, by H type	0.15625
	WSR	5.3 (All rounds) The rate of "L" message in P-Max = that in Neutral, by L type.	0.03125
	WSR	5.4 (Last five rounds) The rate of "L" message in P-Max = that in Neutral, by L type.	0.0625

■ WSR refers to the Wilcoxon signed-rank test. All null hypotheses are two-sided.
■ With six independent sessions, the minimum attainable p-value is 0.03125 (two-sided).
■ P-Max Comp., Neutral Comp., and Uncompromising refer to the P-Max Compromising rule, the Neutral Compromising rule, and the Uncompromising rule, respectively. P-Max and Neutral refer to the P-Max and Neutral Mediators, respectively.
■ For 1.1–2.6, we test for principal-subjects; For 3.1–5.4, we test for subordinate-subjects.
■ In 3.1–3.4, inference labels are abbreviated ('L'='More likely to be L'; 'Same'='Same as prior'; 'H'='More likely to be H').

Appendix B: Supplementary Experimental Results

B.1 Experiment I: Mediator Choice and Inference

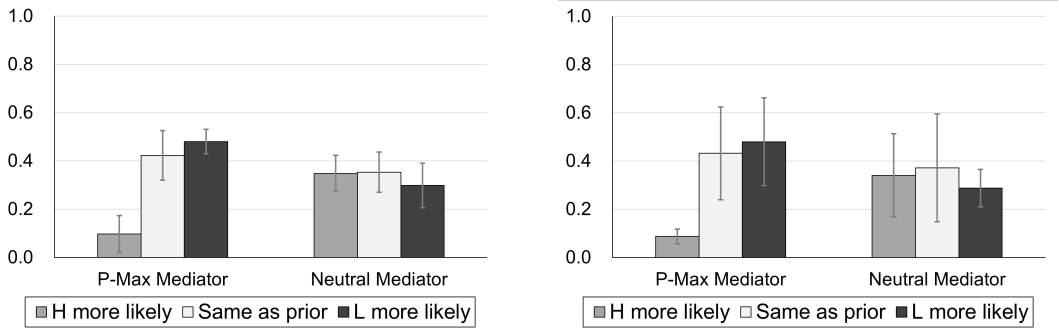
Figure B.1 reports the proportions of mediators chosen by principal-subjects and are played, and those chosen and are declined by subordinate-subjects across all four treatments.



(a) Simple (b) Complex

Figure B.1: Proportion of Mediator Chosen and Played or Declined (Principals)

Figure B.2 reports the frequencies of three inferences by subordinates conditional on either the P-Max or Neutral Mediator chosen by the principal in the Uninformed treatments.



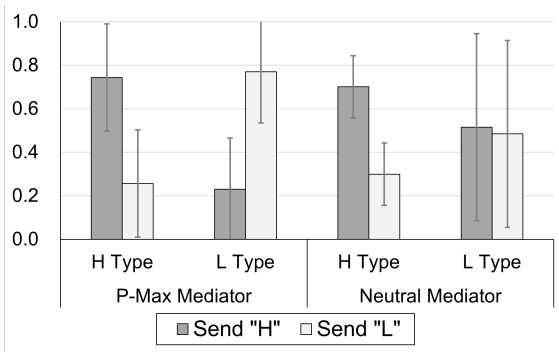
(c) Simple-Uninformed (d) Complex-Uninformed

Figure B.2: Subordinate's Inference Conditional on Mediator Choice (Uninformed)

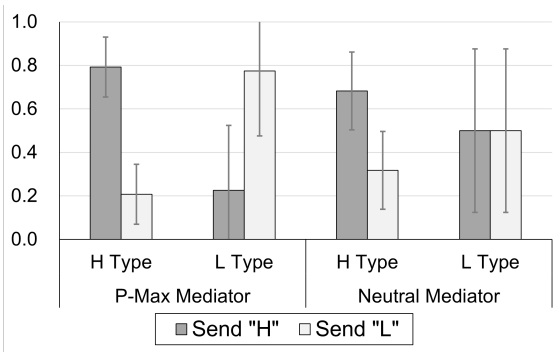
B.2 Principal's Strategy

Figures B.3 reports the frequencies of messages sent by principals by type, conditional on whether the P-Max or Neutral Mediator is chosen, in Part 4 of Experiment I across all

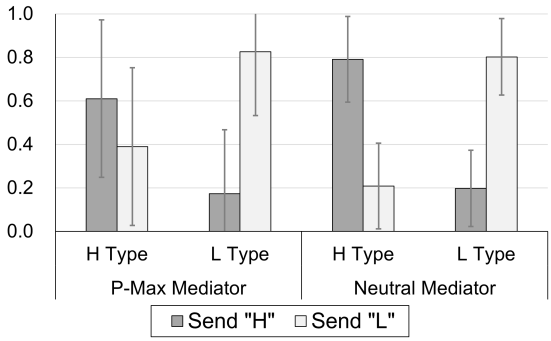
four treatments. Figures B.4 reports the corresponding frequencies in Part 4 of Experiment II. Although full sincerity is never observed, principals send truthful messages with high frequency in both experiments, except in the case of L type principals under the Neutral Mediator in the Simple-Informed and Simple-Uninformed treatments of Experiment I.



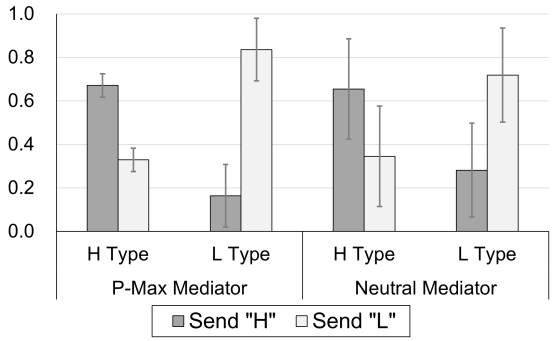
(a) Simple-Informed



(b) Simple-Uninformed



(c) Complex-Informed



(d) Complex-Uninformed

Figure B.3: Principal's Strategy in Chosen Mediator by Type (Part 4, Experiment I)

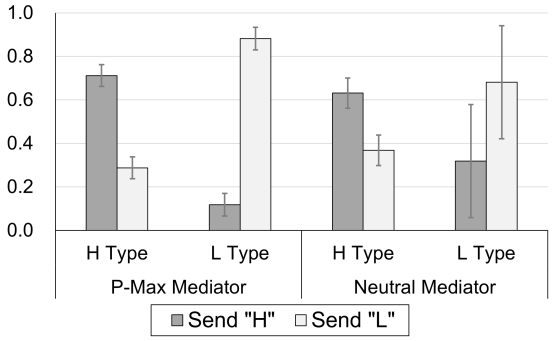
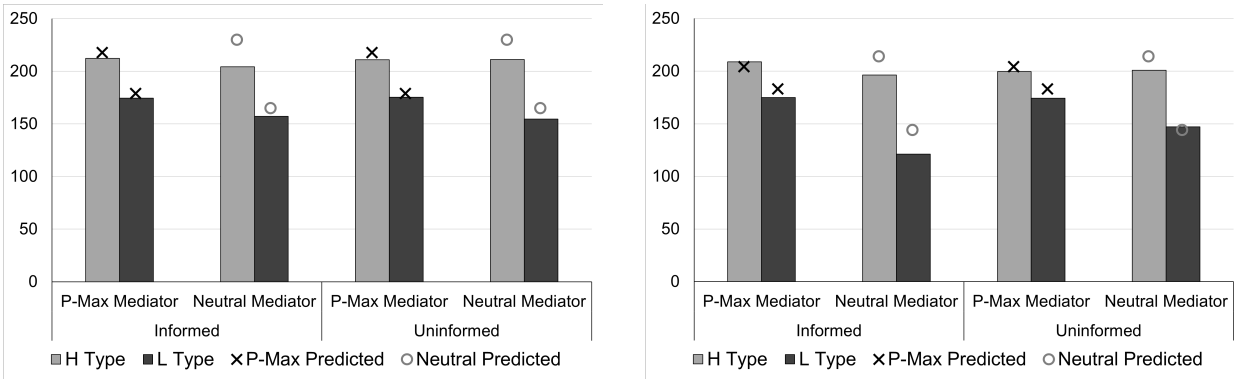


Figure B.4: Principal's Strategy in Chosen Mediator by Type (Part 4, Experiment II)

B.3 Average Payoffs

Figure B.5 reports the average payoffs of each type conditional on the chosen mediator in Part 4 of Experiment I across all four treatments. Figure B.6 reports the corresponding payoffs in Part 4 of Experiment II. The theoretical expected payoffs under the P-Max Mediator are marked with $\times$, and those under the Neutral Mediator are marked with $\circ$.



(a) Simple (b) Complex

Figure B.5: Average Payoff of Each Type (Part 4, Experiment I)

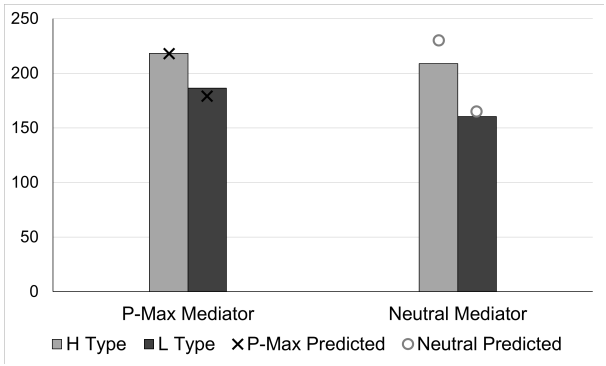


Figure B.6: Average Payoff of Each Type (Part 4, Experiment II)

B.4 Supplemental Analysis: Subordinate's Best Responses

We analyze the mediation subgame (Stage 2) of the mediator selection game under arbitrary interim beliefs held by the subordinate. After the principal chooses and announces a mediator, the subordinate may update his prior belief that the principal is of H type, q, to some posterior q', which we allow to be arbitrary. We do not attempt to fully characterize the sequential equilibria of the game. Rather, our objective is to describe the subordinate-subject's best responses when Stage 2 is played in the lab given his interim beliefs q'. In

doing so, we assume that principals truthfully report their types to their chosen mediators. The subordinate's choices are whether to decline mediation (go to war) and, if he participates, which message to send. Recall that $\theta = 0.75$, $\delta = 0.8$, and that the surplus is set to 400 in the lab. We focus on the case $q = 1/4$; the analysis for $q = 2/5$ proceeds analogously.

First suppose that the principal announces the P-Max Mediator (as used in the lab). The expected payoffs of an H type subordinate under belief q' (a) if he reports truthfully, (b) if he misreports, and (c) if he declines, respectively, are:

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

The payoff from (a) is always strictly greater than that from (b). The payoff from (a) is greater than the payoff from (c) if and only if $q' \geq 16/66 \approx 0.2424$. This threshold differs from 0.25 because the H type's IR constraint does not bind 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, yielding the condition $q' \geq 0.25$. Similarly, the expected payoffs of an L type subordinate under 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 payoff from (c) is always strictly lower than the payoff from either (a) or (b). The payoff from (a) is greater than the payoff from (b) if and only if $q' \leq 5/19 \approx 0.2632$. This threshold differs from 0.25 because the L type's IC constraint does not bind under either the theoretical P-Max Mediator or the version used in the lab.

We treat the two threshold values of 0.2424 and 0.2632 as effectively equivalent to the prior belief $q = 0.25$. With this convention, interim beliefs $q' > 0.25$, $q' = 0.25$, and $q' < 0.25$ correspond, respectively, to the inferences 'more likely to be H than the prior,' 'same as the prior,' and 'more likely to be L than the prior' about the principal's type.

The subordinate's best responses against the principal reporting truthfully in Stage 2 given each inference conditional on the P-Max Mediator chosen are characterized as follows.

- Given 'more likely to be H than the prior' inference, truthful reporting is the best response for the H type, while lying is the best response for the L type.
- Given 'same as the prior' inference, truthful reporting is a best response for either type.

- Given 'more likely to be L than the prior' inference, declining the mediator is the best response for the H type, while truthful reporting 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 an H type subordinate under belief q' (a) if he reports truthfully, (b) if he misreports, and (c) if he declines, respectively, are:

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

The payoff from (a) is always strictly greater than that from (b), and strictly greater than that from (c) except when $q' = 0$ (that is, when the subordinate believes with certainty that the principal is L type). For an L type subordinate, the expected payoffs are:

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

The payoff from (a) is strictly greater than that from (c) except when $q' = 1$, in which case the payoff from (b) is strictly greater than both alternatives. The payoff from (b) is strictly greater than that from (c) except when $q' = 0$, in which case the payoff from (a) is strictly greater than both. The payoff from (a) is greater than that from (b) if and only if $q' \leq 5/19 \approx 0.2632$. Again treating this threshold as equivalent to $q = 0.25$, we can characterize the subordinate's best responses against the principal reporting truthfully in Stage 2 given each inference conditional on the Neutral Mediator chosen as follow.

- Given 'more likely to be H than the prior' inference, truthful reporting is the best response for the H type, while lying is the best response for the L type.
- Given 'same as the prior' inference, truthful reporting is the best response for the H type, and it is a best response for the L type.
- Given 'more likely to be L than the prior' inference, truthful reporting is the best response for either type.

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.

Burdea, Valeria and Jonathan Woon. 2022. "Online Belief Elicitation Methods." *Journal of Economic Psychology* 90:102496.

Casella, Alessandra, Evan Friedman and Manuel Perez Archila. 2025. "On the Fragility of Mediation: Theory and Experimental Evidence." *Journal of the European Economic Association* jvaf049.

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

Danz, David, Lise Vesterlund and Alistair J Wilson. 2022. "Belief Elicitation and Behavioral Incentive Compatibility." *American Economic Review* 112(9):2851–2883.

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.

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

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. 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.