THE UNIVERSITY OF CHICAGO

ESSAYS IN POLITICAL ECONOMY

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE IRVING B. HARRIS
GRADUATE SCHOOL OF PUBLIC POLICY STUDIES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
LIQUN LIU

CHICAGO, ILLINOIS
JUNE 2021

To my parents

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I thank everyone who has walked with me through my journey of graduate studies. I am particularly indebted to the three advisors in my dissertation committee, Scott Ashworth, Ethan Bueno de Mesquita, and Peter Buisseret. Scott has always been my role model for rigorously pushing the boundary of modern political economy; his standard for theoretical research inspires me to write models in an elegant and intellectually serious way. Ethan has always been willing to help; his sharp and insightful comments force me to revisit the foundations of my works and thereby elevate their quality. Peter teaches me to read and write papers in the era of pandemic; his valuable instructions and warm encouragements have deeply shaped the way I conduct and present my academic research.

I am also very grateful to many other professors inside and outside the University of Chicago; their mentoring has greatly expanded my knowledge of research frontiers in economics and political science. They include but are not limited to: Sandeep Baliga, Wiola Dziuda, Georgy Egorov, Yingni Guo, Maggie Penn, Kris Ramsay, Bruno Strulovici, Richard Van Weelden, Georg Vanberg, and Leeat Yariv. For the academic job market, I cannot thank Emerson Niou enough. He made many phone calls to leading Chinese universities on my behalf, and helped me land on an assistant professor position in this unprecedent year.

Graduate school is grueling; I cannot image going through the entire experience without the support and encouragement from many trusted friends. I am particularly grateful to Pablo Balan, Alan Chang, Haohan Chen, Minju Kim, Jialin Li, Lun Li, Yiming Li, Yuan Mei, Marius Ring, Jiafu Wang, Ündes Wen, Di Wu, Yinan Su, Jiming Xu, Renkun Yang, Hanzhe Zhang, Shuang Zhang, and Junlong Zhou. They cheered me up during my blue days; I owe all of them a fancy dinner.

Last but not the least, I am indebted to my parents, Jiusheng Liu and Jie Chen. They brought me up and have always been supportive toward my graduate studies overseas. I dedicate this dissertation to them for their unconditional love.

ABSTRACT

This dissertation consists of three essays examining the agency issues in political economy.

In Chapter 1, I analyze how a careerist delegate carries out reform decisions and implementation under alternative information environments. Regardless of his true policy preference, the delegate seeks retention and tries to signal to a principal that he shares an aligned policy predisposition. Given this pandering incentive, the principal best motivates the delegate's implementation if she can commit to a retention rule that is pivotal on reform outcomes. I characterize an "informativeness condition" under which this retention rule is endogenous, provided that the principal uses an opaque information policy – she observes the delegate's policy choice and outcomes, but not the effort. With other information policies, the principal has to reward congruent policy choices rather than good policy outcomes; her policy interest is damaged by failing to sufficiently motivate reform implementation.

In Chapter 2, I study the limits of mediated conflict resolution when: states have incentives to misrepresent private information about resolve; mediators have limited capacity to enforce agreement; and political leaders bargain in the shadow of audience costs and political bias. With a mechanism design approach, I characterize the conditions under which a mediator may propose a peaceful settlement to resolve the crisis. I find that the availability of peaceful settlements has more to do with political bias than audience costs. The reason is that, regardless of audience costs, the war payoff implied by political bias and the war technology imposes a lower bound that a particular state would ask for from any peaceful settlement. Absent peaceful settlements, I show that some stronger leaders would fight strictly more often and obtain strictly higher payoffs than others. Finally, I examine how the enforcement capacity may impact mediators' ability to implement peaceful settlements.

In Chapter 3, I study a model of testing in which 1) a principal sets a private test standard to infer the type of an agent, and 2) the agent exerts unobserved influence on his/her test result. I characterize conditions under which the principal employs a threshold known as the

"reasonable doubt" to trade off inference errors. Through comparative static analysis, the model offers insights into factors that impact the principal's equilibrium threshold. I find that different test natures, dichotomized by whether an innocent agent outperforms a guilty one when neither exerts influence, may guide comparative static predictions in the opposite directions. The analysis highlights the importance of probing the context and strategic concerns of testing, and generates empirical implications beyond the conventional wisdom that a decision-maker unambiguously leans towards the lesser of two harms.

CHAPTER 1

DELEGATED REFORM DECISIONS WITH CAREER CONCERNS

1.1 Introduction

There is a widespread perception that careerist delegates are timid during policy-making. Career concerns often divert delegates' attention from choosing the most appropriate policy, if there exist alternatives that better impress their principals (e.g. Morris [2001], Canes-Wrone et al. [2001], Maskin and Tirole [2004]). During reform decision-making, in particular, career concerns might distort more than policy choices. Indeed, delegates easily become demotivated at the implementation stage when they see a loomed future for the planned reform.

The following example illustrates this point. Dominic Cummings, a chief figure in the 2016 Brexit campaign and a former senior advisor of Boris Johnson, used to have enormous influence on the British civil servants. In January 2021, several months after stepping down from the Downing Street, Cummings continued his propaganda for the Brexit by posting on the social media "Should I name and shame the senior officials who persistently present a disastrously incorrect picture to ministers?" and "Many will be pleased to know this is about ideologues with EU stars in their eyes, not the virus."[1] To Cummings, supporting the Brexit is a matter of loyalty or congruence for civil servants. Back to the days when Cummings just assumed his position as Johnson's advisor, anecdotes[2] suggest that he pushed this ideology even further. He cued the department aides to support the no-deal Brexit in return for extra money from the Treasury; to department aides, it is "do or die".

1. "Dominic Cummings threatens to expose Remainer civil servants who tried to sabotage Brexit", *Express*, 6 January 2021.

2. "How Dominic Cummings took control in Boris Johnson's first days as Prime Minister". *BuzzfeedNews*. 27 July 2019.

Not all civil servants share Cummings's enthusiasm about Brexit, though. In fact, many become confused, depressed, and/or annoyed at the plan. After long being trapped in the black hole of Brexit, a civil servant complained to *The Guardian*: "Heaven help us if no deal (Brexit) actually happens."[3] Jill Rutter, a former treaury mandarin, put straightforwardly: "Brexit is an article of faith, rather than a pragmatic choice."[4] Others in the Whitehall question the practicability of Cummings's rush and radical changes prior to the Brexit. Dave Penman, the head of the FDA union, pinpointed the source of (de)motivation in reform implementation:" There's a huge difference between bringing in new ideas or radical agendas and implementing untested ideologies which, if they go wrong, will impact upon the delivery of public services to millions of citizens."[5] Observers who have recognized career concerns as real things would naturally ask: how to motivate implementation in the age of reform?

The common solution to the motivation problem is to choose the right kind of transparency. The canonical approach to moral hazard à la Hölmstrom [1979] suggests that in writing a contract, a principal always weakly benefits by observing more about the agent. This wisdom rests on the principal's power to commit to the contract. Absent this power (as is common in the "career concern" literature), knowing more about the agent is not necessarily a blessing. Prat [2005] shows that if the action of a career-minded agent becomes publicly available, then he has incentives to act "congruently" and disregard useful signals; in turn, this hurts the principal's welfare. Ashworth and Bueno de Mesquita [2014] show that constituents may be better off committing not to observe the politicians' policy choice and thereby endogenously choosing a retention rule that motivates the politicians better. While this literature provides many insights about how to mitigate the moral hazard issues in the agency relationship, it remains silent on the motivation problem in a situation where

3. "Many civil servants are depressed – including me. Brexit will do that to you.", *The Guardian*, 26 November 2019.

4. "The civil service must speak truth to Boris (and his Cabinet)". *The Institute for Government*. 25 July 2019.

5. "Dominic Cummings role provokes alarm inside civil service". *The Guardian*, July 25 2019.

pandering to the congruent policy is inevitable.

This paper seeks to address the motivation issues in reform policy-making with a formal model. I build on Hirsch [2016] and consider a two-dimensional policy-making environment consisting of policy choices and implementation. The main departure from Hirsch [2016] is that the person-in-charge has reputation concerns instead of heterogeneous beliefs from a principal. As such, I consider how the principal employs information policies rather than delegating the control right of projects to motivate the agent. I ask: how would different information policies incentivize different patterns of policy-making and implementation? Which information policy would induce the optimal policy consequences?

I study a setting in which there is a policy-driven principal and a career-minded agent. They face the policy choice between a reform and the status quo. The status quo is safe: it confers a state-independent payoff to both players. The reform is risky: it may fail either because the reform by nature is a bad decision (i.e. the reform is not "appropriate"), or because it is not well executed. The principal prefers a successful reform to the status quo to a failed reform. However, she lacks necessary information expertise and/or implementation capacity to carry out a reform well. She delegates the policy-making and implementation to an agent with the hope of utilizing his expertise.

In this delegation, there are two agency issues à la Canes-Wrone et al. [2001]. One is that the principal is unsure whether the agent has an aligned policy predisposition. Some open-minded agents may share the principal's policy preference. Others may be politically conservative; they are hostile to any reform attempt, and strictly prefer the status quo no matter what. I label two types "congruent" and "noncongruent" respectively (see also Morris [2001], Fox [2007]). The other agency issue comes from the agent's reputation consideration during his policy-making. He may value a congruent public image for human reasons. For example, a bureaucratic may enjoy appearing receptive and flexible to changes in the public domains. Another perhaps more important reason is that the agent's reputation may affect

his career – in policy-making, subordinates want to stay in tune with their superiors (as the Cummings example suggests). Either way, the agent may disregard his private information and pander to the "popular" or "congruent" actions.

The goal of the principal is to further her policy interests by choosing the optimal policy-making information regime. She may require the agent to disclose his 1) policy choice, 2) policy choice, outcome, or 3) policy choice, outcome, and efforts. Label these regimes as "observable policy", "observable outcome", and "full transparency"; they all describe relevant reform decision-making environments. The first two regimes are standard in the literature (e.g. Prat [2005] and Fox [2007]). For example, a representative voter may or may not observe the incumbent politician's policy consequences before the election day. The third one seems a little unconventional: we often attribute an agent's effort to "moral hazard" that is unobserved by the principal (e.g. Hirsch [2016]). Per Hölmstrom [1979], we may interpret this effort-observability as a reduce-form representation that a principal may receive signals about congruence in addition to policy choices and outcomes. Alternatively, it describes a situation in which the principle does not keep the agent at the arm's length. This regime is most plausible when the principal may strictly discipline an agent and *want* to know the reason why a project fails (see also Crémer [1995]). For example, a university council may adopt various information policies to determine whether a junior faculty has been actively engaged in innovative research. In addition to research agenda and output (publication record), the council may use the number of research grants to proxy this person's effort. I examine how different information policies shape the agent's behavior and feedback to the principal's policy payoff.

Main results. I show that, for a large set of parameters, the principal is best off with an "opaque" information regime – she observes the agent's policy choice and outcome, but not his implementation effort. To fix ideas, for this moment let us suppose that the principal is quite certain about facing a congruent agent; and, a congruent agent always panders to

the reformist policy. These two hypotheses reduce the principal's problem to motivating the congruent agent to reform well. The opaque information regime may prevail because the principal can promise to retain conditional on a successful reform. Foreseeing that his future career is tied to reform outcomes, the agent works very hard. But the principal's promise does not come easy; it suffers from the lack of commitment (as is common in the prospective voting literature). To lend credibility to this retention strategy, I identify an "informativeness condition" under which the principal draws positive (negative) inference about the agent's congruence from a successful (unsuccessful) reform. In the opaque information regime, this condition insures the principal to commit to this pivotal retention rule, and enables her to elicit the maximum possible efforts from a congruent agent.

The agent becomes less motivated if the principal has assess to a poorer or richer set of information about policy-making. Suppose the principal observes the policy choice only. In this case, a congruent agent always secures office by posturing at the reformist policy. The reason is that, when the principal observes nothing but the policy choice, and if she is convinced that a congruent type always reforms, then initiating the reform cannot be bad news about congruence. Since reform outcomes are not pivotal on the retention decision, the congruent agent exerts less effort than what he does in the opaque regime. Similarly, in a fully transparent regime a congruent agent does not worry too much about whether reform failure would cost his future career. The logic is slightly different though: the congruent agent can secure office by signaling his congruence in the dimensions of policy choices and efforts. The principal would strictly prefer the opaque regime to this fully transparent one, if the effort for separation is lower than the effort motivated in the opaque regime.

Now we are in a position to justify why a congruent agent always panders to the reformist policy across all three information regimes. The rationale is that, since the noncongruent agent does not derive any policy benefit from the reformist policy, even an "unlucky" (observing a bad signal) congruent agent is more keen on reform than a "lucky" noncongruent

one. As such, keeping the status quo policy always makes the principal suspicious that she is facing a noncongruent type. To convince the principal of his congruence, the congruent agent must initiate a reform even during bad times.

The motivation effect of opaqueness brings a new perspective to the literature of transparency. A near-axiom is that there is a sorting (select the right agent) and discipline (motivate agent's effort) tradeoff of transparency (e.g. Holmström [1999], Dewatripont et al. [1999], Crémer [1995], Fox [2007]). The wisdom partially applies to the model considered in this paper: as the principal observes more about the agent's decision-making, a noncongruent agent's incentive to mimic a congruent type diminishes. Specifically, two-type agents pool on the reformist policy when the principal only observes the policy choice; they separate by choosing their favorite policies when the principal observes everything. But the discipline effect is nonmonotone with respect to the transparency of decision-making. The agent is most motivated to implement well when he gambles his career on reform outcomes. To make this event pivotal in her retention decision, the principal must leave some opaqueness in the information policy.

Finally, I provide several comparative static results to unveil the mechanics of the principal's optimal information policy. The principal essentially chooses between the observable outcome regime and full transparency; the observable policy regime is strictly dominated because pandering creates pure distortion there. I start with an environment in which the the observable outcome regime is optimal, and assess whether the principal would like to switch policies after parameter changes. The main lesson is that, the observable outcome regime would continue to prevail if parameter changes either 1) strengthen the motivation effect inherited in the observable regime, or 2) hinder the noncongruent agent to mimic the congruent type. In both situations, the observable outcome regime reinforces its superiority by better motivating the agent than full transparency. More interesting is the result with respect to office rent. A larger rent better motivates the agent in the observable outcome

regime; it also asks more effort from a congruent agent to separate under full transparency. Either force may dominate depending on model parameters.

This paper proceeds as follows. Section 1.2 presents the delegation game. Section 1.3 characterizes the equilibria subject to different information policies. I identify a parameter restriction labeled the "informativeness condition", and then prove how this condition makes the retention decision pivotal on reform outcomes. Section 1.4 examines the welfare implications of different information policies. Section 1.5 concludes.

1.2 Model

Setup. To make the main point, consider a simple one-period model of policy choice and implementation. There is a reform committee consisting of a principal and an agent. The agent is delegated the task of choosing a policy x from a binary set $X = \{r, q\}$. r stands for a "reform" policy, and q stands for the "status quo" policy.

The reform is risky: its outcome y could be either success ($y = 1$) or failure ($y = 0$). Following Hirsch [2016], I assume that "choosing well" and "implementing well" are complementary in achieving reform success. Specifically, a reform could be good ($\omega = G$) or bad ($\omega = B$) with the common prior $P(\omega = G) = \phi \in (0, 1)$. A bad reform always fails; a good reform succeeds with a probability equal to the agent's implementation effort $e \in [0, 1]$. The principal does not know the nature of the reform. The agent receives a signal $s \in \{g, b\}$ of accuracy p about ω i.e. $P(s = g|\omega = G) = P(s = b|\omega = B) = p$. Assume that $p \geq \frac{1}{2}$.

The status quo policy is safe: it confers a constant payoff to players. Furthermore, it does not require any implementation effort. Notate $y = \emptyset$ if the agent takes the status quo policy; it says the reform is never chosen and bears no consequence.

The agent has a private type $t \in T := \{c, n\}$; he may be congruent ($t = c$) or non-congruent ($t = n$). The principal does not observe t; she assigns a prior probability $P(t = c) = \pi \in (0, 1)$. A congruent agent shares the policy preference with the princi-

pal: they both value reform success at 1, the status quo at $d \in (0,1)$, and reform failure at 0. Formally, we write their policy payoff function as $v(x,y)$, where $v(r,1) = 1, v(r,0) = 0$ and $v(q,\emptyset) = d$. A noncongruent agent also values the status quo at d, but he values any reform outcome at 0. His policy payoff function can be summarized as $v_n(x,y) = d \cdot 1\{x = q\}$. The definition of congruence suggests that both types of agent prefer the status quo policy to a bad reform; however, they disagree with respect to whether a good reform is valuable. Substantively, we may interpret (non)congruence as describing the agent's policy preference. For example, there may be agents who are captured by interest groups and do not derive benefit from policy changes [Fox, 2007]. Alternatively, we may interpret congruence in terms of which policy legacy the agent wants to leave (see also Maskin and Tirole [2004]). A political conservative may not want a reformist policy to appear on her/his curriculum at all.

The total payoff to the principal depends on the policy consequence; her preference for retaining a congruent agent is lexicographic[6]. In line with Fox [2007], this assumption suggests that the principal's primary concern is policy-making. If the principal also attaches a nontrivial weight to selecting the right agent, then all else equal she would lean toward a decision-making environment that enables her to learn more about the agent (see also Fox and Van Weelden [2012]). To simplify matters, I assume that the principal retains this agent whenever she is indifferent between retention and replacement. Formally, let I be the information set on which the principal conditions for her retention decision $D \in \{0,1\}$; she wants to retain the agent ($D = 1$) whenever her posterior belief of the agent being congruent, $P(t = c|I)$, is at least π; she replaces ($D = 0$) otherwise.

The total payoff to the agent depends on the policy consequence, retention decision, and the implementation cost. The agent receives an office rent $R > 0$ whenever $D = 1$. For each effort level $e \in [0,1]$ the agent bears an implementation cost $\frac{e^2}{2\lambda}$. Here λ parameterizes the

6. We may micro-found this assumption by adding an $\epsilon > 0$ weight of retention, and focusing on the limiting case $\epsilon \downarrow 0$. See Appendix 1.6.5 for a detailed discussion.

agent's cost sensitivity: a higher λ means that the agent bears a lower cost of implementation.

Here is the summary of payoff.[7] The principal's utility is $u_p(x, y) = v(x, y)$; a congruent agent's utility is $u_c(x, y) = v(x, y) - \frac{e^2}{2\lambda} + R \cdot D$; a noncongruent agent's utility is $u_n(x, y) = v_n(x, y) - \frac{e^2}{2\lambda} + R \cdot D$.

Information environment. During the delegated reform decision-making, the agent always reports his policy choice x to the principal. Whether the principal may gather extra information and thereby better discipline the agent depends on her involvement in the reform. This renders the follow three information regimes most relevant: observable policy, observable outcome, and full transparency. With respect to three regimes, the principal's information set I contains policy choice only ($I = \{x\}$); policy choice x & outcome y ($I = \{x, y\}$); and policy choice x, the effort e & the outcome y ($I = \{x, e, y\}$).

Remark 1. *The observable outcome regime gets its name because the policy outcome y is a sufficient statistic for the policy choice x. To see this, the principal can always invert $x \in \{r, q\}$ from any outcome $y \in \{1, 0, \emptyset\}$. This perfect invertibility prevents the observation in Prat [2005] that the principal may be better off committing not to observe the action but solely the outcome of policies.*

Sequence of moves. The game moves as follows:

1. Nature draws ω, s, and t.

2. The agent observes $s \in \{g, b\}$.

3. The agent chooses $x \in \{r, q\}$ and effort $e \in [0, 1]$.

4. Nature determines the outcome of the project y; the policy payoff realizes.

7. For presentation clarity, we omit the dependence of y on e.

5. The principal decides on whether to retain the agent according to different information environments I.

Assumptions. Notate $\mu_+ = P(\omega = G|s = g) = \frac{\phi p}{\phi p+(1-\phi)(1-p)}$ and $\mu_- = P(\omega = G|s = b) = \frac{\phi(1-p)}{\phi(1-p)+(1-\phi)p}$. μ_+ and μ_- stand for the posterior beliefs that the reform is by nature good after good ($s = g$) and bad ($s = b$) signals.

Assumption 1 (Useful signals)**.** $\mu_+ > \sqrt{\frac{2d}{\lambda}} > \mu_-$.

This condition suggests that the signals are "powerful" or very informative: the signal s can sufficiently sway one's opinion with respect to whether the reform is good or bad.

Assumption 2 (Moderate office rent)**.** $\min\{(1 + R)\mu_-, R\mu_+\} > \sqrt{\frac{2d}{\lambda}} > R\mu_-$.

This condition says that the office rent is moderate. It makes the problem interesting. Otherwise, very large office rents drive all types of agents to pander to the congruent action r; very small office rents do not discipline the agent's action at all.

I also impose $\lambda(1 + R) \leq 1$ to ensure that $e \in [0, 1]$. In the appendix Lemma 4, I show that this parameter restriction together with Assumption 2 implies that $R > 2d$.

Solution concept. In the signaling game, the agent chooses an action $a \in \{r, q\} \times [0, 1]$ based on his type $t \in \{c, n\}$ and signal $s \in \{g, b\}$. His (mixed) strategy $\sigma(\cdot|t, s)$ specifies a probability distribution over actions a for each type-signal tuple (t, s). The principal decides whether to retain this agent based on her information set I; her (mixed) strategy is $\sigma_p : I \to [0, 1]$.

For each information regime, I look for the existence of Perfect Bayesian Equilibrium (PBE) in pure strategy. The equilibrium condition dictates that no players admit any profitable deviation from the equilibrium strategy; and, players form beliefs by the Bayes' rule whenever possible.

A preliminary observation is that the game may admit numerous equilibria supported by ad hoc off-path beliefs. Take the case of the observable policy regime. We may specify a pooling equilibrium in which both types of agent choose $x = q$ regardless of signals, and the principal holds the off-path belief that $P(t = c|x = r) = 0$. This equilibrium can be sustained if career concerns outweigh policy considerations; it is nonetheless not satisfying – the congruent agent may make a public speech to convince the principal that that he is willing to take action $x = r$ when the signal is good. The principal might buy this argument, for indeed only a congruent agent may benefit from this kind of action.

More complications arise when we talk about the "type" of the agent. For the principal's retention strategy, what matters is the agent's congruence; but for the agent who decides on the policy choice and implementation, the right kind of "type" definition must also account for the agent's signal. These complications call for an extended "type space" beyond the congruence/noncongruence dictonomy; more importantly, they ask for a powerful refinement to restrict off-path beliefs that deems a PBE reasonable.

I apply the universal divinity refinement from Banks and Sobel [1987] to the PBEs of this game. In this game, an agent's type is summarized in a tuple (t, s) where t is his payoff type and s is his signal. Upon any deviation, this refinement attaches more weight to the type who may benefit more than other types relative to his equilibrium outcome. Since only the agent's payoff type matters in the principal's retention, I label an agent "congruent/noncongruent" to mean that this agent has a payoff type c/n.

The university divinity refinement assigns the following beliefs off the equilibrium path:

Definition 1 (Strong off-path beliefs)**.** *A PBE is supported by strong off-path beliefs whenever*

1. *If $x = q$ is off the path of play for both types of agent, then the principal assigns probability 1 to the event that the agent is noncongruent.*

2. *If $x = r$ is off the path of play for both types of agent, then the principal assigns*

probability 1 to the event that the agent is congruent.

I justify this definition in the Claim 1 of the appendix. Similar to Fox and Jordan [2011], the strong off-path beliefs attach probability 1 to whichever payoff type that may benefit more by deviating from the equilibrium strategy. They are particularly handy in contemplating the principal's equilibrium retention strategy.

From now on, I label any PBE surviving this refinement as an "equilibrium".

1.3 Policy-making under alternative information environments

In this section, I characterize the agent's equilibrium behavior under different information regimes. I examine how transparency – the degree to which the principal may observe the policy-making process – distorts the agent's behavior that may or may not benefit the principal.

It is useful to establish a benchmark where the agent has no career concern. This situation is plausible if either the agent is a lame duck or if the principal observes absolutely nothing about the agent's behavior.

Fact. *Absent career concerns, a congruent agent chooses r with effort $\lambda\mu_+$ after $s = g$, and chooses q after $s = b$. The noncongruent agent always chooses q.*

All omitted proofs are relegated to the appendix.

The observation is straightforward: the agent simply chooses his favorite policy. A congruent one reforms whenever receiving a good signal; a noncongruent one always keeps the status quo.

1.3.1 Observable policy

First consider the case in which the principal observes the policy only. This environment echoes Fox [2007]. Unsurprisingly, making policy choices observable to the principal would

incentivize a career-minded agent to pander to the congruent action; it may have a deleterious effect on the welfare of the principal.

Result 1. *Under the observable policy regime, there exists a unique equilibrium in pure strategy. In this equilibrium, the congruent agent always chooses r; he implements with effort $\lambda\mu_+$ after $s = g$, and $\lambda\mu_-$ after $s = b$. The noncongruent agent always chooses r with effort 0 regardless of signals. The principal always retains the agent.*

When the principal observes the policy choice but not the outcome, she retains on whether the agent has taken the congruent action. Since the office concern dominates the payoff of status quo policy ($R > 2d$), the agent initiates the reform regardless of signals in order not to appear noncongruent. But this means that the implementation quality of reform is very low: the principal would rather keep the status quo than leave a reform in the hand of a noncongruent agent. The observation that pandering incentives may hurt the principal is in tune with Fox [2007].

1.3.2 Observable outcome

Now let us improve the principal's involvement in the policy-making. Suppose the principal observes the outcome $y \in \{0, 1, \emptyset\}$ before making her retention decision. She understands whether agent has succeeded in a reform $(r, 1)$, failed in a reform $(r, 0)$, or kept the status quo $(q, \emptyset)$. Her set of retention strategy is rich. For example, congruent behavior ($x = r$), successful implementation ($y = 1$), no failure ($y \neq 0$) are all sensible retention criteria.

To simplify matters, I consider when the reform outcomes convey retention-relevant information. Note that the congruent agent is always more motivated to work on reforms. He initiates a reform with a higher probability; and, conditional on a reform, he exerts more effort than a noncongruent agent. Taken together, reform success is always good news about the agent's congruence. But reform failure is a different story. *A priori*, the principal cannot

determine whether reform failure is more likely to come from the noncongruent type. This is because the congruent agent fails at a lower probability but he reforms more often.

Below I provide a sufficient condition that guarantees reform failure to be bad news about congruence. Notate $\gamma = \frac{1-p}{p}$ and $z = \frac{1-\phi}{\phi}$:

Definition 2 (Informativeness condition)**.** $z - \lambda\frac{1}{1+\gamma z} \leq \gamma[\lambda(1+R)\frac{\gamma}{\gamma+z} - 1]$.

The informativeness condition derives from the following thought experiment: for reform outcomes to be pivotal in retention, it must be that both types of the agent reform on path[8]. If both types reform whenever observing the same signal(s), then reform outcomes must be pivotal because failure suggests a lack of motivation. More interesting is the case in which a congruent type always reforms while the noncongruent agent reforms only if the signal is good. As before, a successful reform must be good news about congruence; but now a failed reform might not be bad news unless a lucky noncongruent agent becomes sufficiently demotivated. From this conjectured strategy profile we compute the informativeness condition.

The informativeness condition depends on the following parameters: the signal accuracy p, the prior probability of a good state ϕ, the implementation cost parameter λ and the office rent R. p and ϕ pin down how *more often* a congruent type reforms relative to a noncongruent type (i.e. receiving a bad signal). If p or ϕ increases, a bad signal becomes less likely to occur; therefore, the principal would learn from a reform decision that the agent probably has received a good signal. Consequently, any reform failure is more likely to come from bad execution rather than bad policy choices. This makes reform failure really bad news about congruence, because it reveals the agent's lack of motivation. λ and R describe how *more motivated* a congruent type becomes than a noncongruent type in a reform. If λ or R becomes larger, the congruent agent has very strong motivation to reform well; hence,

8. Otherwise, either they pool on the status quo policy regardless of signals, in which case the action of reform in itself is good news; or one type reforms and the other does not, in which case the principal learns perfectly who reforms.

the principal should attribute reform failure to the lack of policy motivation, and deduce that the agent is probably noncongruent.

The following lemma summarizes these observations.

Lemma 1. *Let $v = (\lambda, R, \phi, p)$ and $v' = (\lambda', R', \phi', p')$ be two vectors of parameters where $v' \geq v$ component-wise with at least one inequality strict. If the informative condition holds for v, then it also holds for all v'.*

We are ready to state the main result of this subsection:

Result 2. *Suppose the informativeness condition holds. Then there exists a unique equilibrium in pure strategy. In this equilibrium, 1) the congruent agent always chooses r. He implements with effort $\lambda(1+R)\mu_+$ after $s = g$ and $\lambda(1+R)\mu_-$ after $s = b$. 2) the noncongruent agent chooses r with effort $\lambda R\mu_+$ after $s = g$, and $x = q$ after $s = b$. The principal retains the agent after $(r, 1)$ and replaces after $(r, 0)$ or $(q, \emptyset)$.*

The equilibrium has several interesting features. The noncongruent agent's project choice is always responsive to signals; the congruent agent always initiates the reform even if the signal is bad. At face value, their equilibrium behaviors seem to suggest that the noncongruent agent appears more accountable: due to career concerns, the congruent agent is too "timid" to hold on to his policy judgement. That said, the noncongruent agent's "good judgement" does not come from his good policy intention; instead, he gambles for policy success to secure office. This induces an excessively high risk of reform failure.

It is also worth mentioning that the principal selects on both "good choice" and "good execution". This selection rule is not only empirically plausible, but also complements the theoretical models of transparency. Earlier works such as Fox and Shotts [2009] have acknowledged that a principal may reward an agent on the basis of congruent actions or good execution (but not both). My result suggests that good choices and execution may be complementary in proving congruence.

Remark 2. *If the office rent is too large i.e.* $R\mu_- > \sqrt{\frac{2d}{\lambda}}$*, then there exists an equilibrium in which the agent always chooses* $x = r$*. After* $s = g$*, the congruent agent exerts effort* $\lambda(1+R)\mu_+$ *while the noncongruent agent exerts effort* $\lambda R\mu_+$*; after* $s = b$*, the congruent agent exerts effort* $\lambda(1 + R)\mu_-$ *while the noncongruent agent exerts effort* $\lambda R\mu_-$*. The principal retains if she observes* $(r, 1)$*; she replaces if she observes* $(r, 0)$ *or* $(q, \emptyset)$*. This equilibrium describes another possibility that the reform outcome might be pivotal in retention. Compared with the main model, here the principal suffers from a demotivated (noncongruent) agent reforming at a bad time.*

1.3.3 Full transparency

We now turn to examining the consequences of a fully transparent decision-making environment. Suppose the principal also observes the agent's effort in addition to his policy choice and outcome. Now the game resembles the Spence model (e.g. Spence [1973]), with the only difference being the agent signaling on the dimensions of both the policy choice and effort. Full transparency empowers the principal to discipline the agent to the maximal level, but it also incentivizes a noncongruent agent to act to his own interest.

As a signaling model, this game inevitably exhibits equilibrium multiplicity. Among the class of separating equilibria, the divinity refinement uniquely selects the least costly one also known as the "Riley outcome" (Riley [1979]); for some parameter values, the model also admits a continuum of pooling equilibria. A detailed discussion is relegated to the appendix Lemma 6.

I focus on separating equilibria for two reasons. Substantively, selecting on the separating equilibria reflects the wisdom that more information tends to improve sorting. Technically, the separating equilibrium is more robust to parameter changes. To see it, the existence of a separating equilibrium is guaranteed because a congruent agent is a better-motivated reformer (i.e. the single-crossing property). The existence of a pooling equilibrium is a

different story. In Lemma 6, I show that a pooling equilibrium surviving the universal divinity refinement exists only if the cost of implementation is not too low; otherwise, a congruent agent exerts so much effort that a noncongruent type finds it not worthwhile to match.

Result 3. *Under full transparency, there exists a unique separating equilibrium that survives the universal divinity refinement. In this equilibrium the congruent type always chooses r; he implements with effort $e_H = \max\{\sqrt{2\lambda(R-d)}, \lambda\mu_+\}$ after $s = g$ and $e_L = \max\{\sqrt{2\lambda(R-d)}, \lambda\mu_-\}$ after $s = b$; the noncongruent type always chooses q. The principal's belief about the type $(t, s) \in \{c, n\} \times \{g, b\}$ is that $P((c, g)|(r, e_H)) = 1$, $P((c, b)|(r, e_L)) = 1$ and $P((c, s)|(q, \emptyset)) = 0$ for all $s \in \{g, b\}$; the off-path belief is $P((c, s)|(r, e) : e < e_H, e \neq e_L) = 0$ for all $s \in \{g, b\}$ and $P((c, g)|(r, e) : e > e_H) = 1$. She retains whenever observing (r, e) with $e \geq e_H$ or (r, e_L).*

Since a congruent agent values reform success more than the noncongruent one, a fully separating equilibrium always exists. In this equilibrium, the congruent agent takes the congruent policy choice $x = r$, and exerts enough efforts to separate from the noncongruent type.

Full transparency creates two effects. First, the policy-making has become purely "partisan": the congruent agent always reforms, and the noncongruent one always take the status quo. This improves sorting (the congruent type) but hurts discipline. Second, the office rent affects the agent's behavior via the effort to separate. If the rent is large enough, the principal may prefer the congruent agent to initiate the reform even when the signal is bad.

1.4 Welfare comparison

In this section, I examine the role and consequences of transparency in reform policy-making.

1.4.1 Career concerns and transparency

Let us compare Results 1-3 with the no career concern benchmark. The following patterns are in line with previous literature: as the principal becomes more involved in the policy-making, she can better discipline the agent with office rent (see also Crémer [1995], Dewatripont et al. [1999]); but then the agent also panders more often (Canes-Wrone et al. [2001], Maskin and Tirole [2004], Fox [2007]). More important are the following novelties:

First, the sorting-discipline patterns are not monotone in the level of transparency. On the one hand, the noncongruent agent finds it harder to hide himself when the principal obtains more information about decision making (monotonic sorting). On the other hand, the congruent agent always panders to the congruent action; but his motivation to *reform well* is strongest when the transparency level is intermediate (non-monotone discipline).

Second, career concerns entail a motivation effect via a retention rule that is pivotal on reform outcomes. Since a successful reform brings the double benefits of retention and policy, the agent works very hard with the hope of impressing the principal with a good reform outcome. This equilibrium behavior illustrates how motivating *despite* pandering is possible.

Third, this motivation effect may happen only under an intermediate level of transparency. The reason is that, a pivotal retention rule requires the principal to retain on a "noisy" policy outcome – she never perfectly learns the congruence of the agent from it. But the principal's desire to learn "something" from policy outcomes is against the spirit of full opaqueness or full transparency; they often end up with no learning or perfect learning.

1.4.2 Optimal transparency

We are in a good position to assess the welfare ranking of different information environment. I restrict attention to the equilibria described in Results 1-3, and measure the welfare according to the (principal's) policy payoff. Note that Result 2 describes an equilibrium unique among

others: else, the principal retains on "congruent behaviors".

Proposition 1. *The observable policy regime induces the lowest welfare. There exists an open set of parameter values* $w = (p, \phi, d, \lambda, R, \pi) \in \Omega$ *such that the observable outcome regime induces the highest welfare.*

The core of this proposition concerns which information regime elicits most efforts from an agent who panders to reform. The observable policy regime is dominated because the principal is hurt by unproductive pandering. There, an agent on average takes the congruent policy (reform) without inputting serious effort.

By contrast, the observable outcome regime motivates the agent well by linking the retention decision with reform outcomes. Here, all congruent agents and a good-luck non-congruent agent work very hard to gamble their fate on reform success. While sometimes the principal might find the timing of reform suboptimal (i.e. congruent agent with a bad signal & noncongruent agent with a good signal), she can more or less rest assured that the agent would input a reasonable amount of effort for his better tomorrow. As such, this regime may dominate if its motivation effect is strong enough.

The full transparency regime also avoids unproductive pandering. It elicit efforts from a congruent agent who wants to distinguish himself from a noncongruent one; it also encourages the noncongruent agent to stay with the status quo policy. The principal benefits more from this information policy as the bar for separation becomes higher; this leaves open the possibility that full transparency might prevail.

1.4.3 Comparative Statics

To shed light on the determinants of the principal's optimal information policy, in this subsection I provide several comparative static results.

To bypass the myriad of parameter restrictions, I start with an environment in which the observable outcome regime is optimal. After that, I examine what/how parameter changes

might shift the principal's optimal information policy. When performing the analysis, I leave Assumptions 1-2 and the informativeness condition intact; in doing so, players would use the same equilibrium strategy before and after the exogenous environment changes. This makes comparative statics as transparent as possible.

Here is the main result:

Proposition 2. *Given a vector $w = (p, \phi, d, \lambda, R, \pi)$ and suppose the observable outcome regime is the principal's optimal information policy under w. Then*

1. *The principal would continue to use this regime if ϕ increases; moreover, the principal strictly benefits from this.*

2. *Suppose further that $2(R-d) < \lambda$. Then 1) there exists $p' \in (\frac{1}{2}, 1)$ such that for all $p \in (p', 1]$, the principal would continue to use this regime if λ increases; 2) $\exists p'' \in (\frac{1}{2}, 1)$ such that for all $p \in (p'', 1]$, the principal would continue to use this regime if p increases. In both cases, the principal strictly benefits from this parameter changes.*

3. *An increase in R may cause the principal to switch to the full transparency regime. Specifically, fixing sufficiently high p, ϕ, and π, and assuming $\lambda(1+d) \leq \frac{1}{2}$, there exists a pair $\underline{R} = \underline{R}(\lambda, d)$ and $\bar{R} = \bar{R}(\lambda, d)$ such that 1) for all $R \in (\underline{R}, \bar{R})$ the full transparency regime dominates; for $R > \bar{R}$ and $R < \underline{R}$ the observable outcome regime dominates. 2) $\underline{R}$ is increasing in λ and d; $\bar{R}$ is decreasing in λ and d.*

This proposition encapsulates the following idea: for the principal to maintain the observable outcome regime after parameter changes, she must benefit more from (both types of) the agent's incentive to gamble for resurrection than a congruent agent's incentive to separate.

A larger ϕ (stronger prior of a good state) makes the observable outcome regime more appealing to the principal. The intuition is that, with a stronger prior the agent is more certain that the reform is good by nature. Consequently, in the observable outcome regime

he works harder than before. This "strong prior" effect extends to the full transparency regime only if 1) the agent is congruent and 2) $\sqrt{2\lambda(R-d)} < \lambda\mu$ for some $\mu \in \{\mu_+, \mu_-\}$; otherwise, the agent's policy choice and implementation effort remain invariant to the prior. But under these two conditions, the observable outcome regime strictly dominates because the agent exerts more effort there ($\lambda(1+R)\mu > \lambda\mu$). Taken together, a stronger prior ϕ always biases the principal's optimal information policy in favor of the observable outcome regime.

Under some circumstances, the principal would maintain the observable outcome regime when λ and/or p increase. The following observation is key: if the bar for separation is low ($\sqrt{2\lambda(R-d)} < \lambda\mu$), then the congruent agent must exert less effort under full transparency than in the observable outcome regime. Sufficiently accurate signals guarantee this low bar for separation. To see it, high accuracy p does two things: subject to $2(R-d) < \lambda$, it insures easy separation when the signal is good by inducing a sufficiently high posterior μ_+ ($\sqrt{2\lambda(R-d)} < \lambda\mu_+$); it also allows the principal to do welfare calculus conditional on a good signal realization[9]. Since a higher λ and/or p only reinforce easy separation, the principal would find it worthwhile to maintain the observable outcome regime.

Matters are different if the office rent R increases. The rationale is straightforward: more attractive office better motivates an agent in the observable outcome regime, but it also incentivizes a congruent agent to work harder for separation. *A priori*, it is hard to tell which effect dominates. In the appendix, I show that the relative magnitude of the motivation and separation effects may go either way.

9. With a bad signal, the reform is doomed to fail and confers almost zero payoff; the status quo policy confers $d > 0$. Since the agent's equilibrium strategies remain invariant to the parameter changes, the welfare impact conditional on a bad signal is minimal from the principal's perspective

1.5 Conclusion

This paper examines how a principal may use information policies to motivate a career-minded delegate in the context of reform policy-making. Key to this motivation problem concerns whether a principal may credibly commit to a retention rule that is pivotal on reform outcomes. I identify an informativeness condition under which this retention rule is credible, provided that the principal adopts an opaque information policy – she observes the delegate's policy choice and outcomes, but not the implementation effort. I show that, for a nontrivial set of parameters, indeed this opaque information regime makes the principal best off. With other information policies, this retention rule is not credible because the principal would in equilibrium learn too much or too little. She would retain according to the delegate's reform decisions; her policy interest is damaged by failing to sufficiently motivate the delegate to reform well.

How should we interpret the results substantively? They provide an informational remedy to the principal's motivation problem, when there lack formal institutions to credibly reward successful reforms. In the contract theory textbooks, the motivation issue is easily resolved if the principal can commit to a monetary reward equivalent to the office rent for successful reforms. But this sort of commitment is not always feasible in the real world. More often than not, monetary compensations are associated with bribery or corruption. It is hard to image that Cummings might lure fellows in the Whitehall with bonus linked to a successful no deal Brexit. Absent the right incentive, however, reformist policies might bear excessively high risks due to the delegate's underprovision of effort. We have encountered numerous cases in which reform failure begets disastrous consequences. Per this paper, perhaps the principal benefits by maintaining an arm's length relationship with the agent; other monitoring and auditing devices are inefficient even if they are technologically feasible.

1.6 Appendix

1.6.1 Preliminary

Definition 3. *An arbitrary event $I \in \mathcal{I}$ is* **neutral** *news about congruence if $P(t = c|I) = \pi$; it is* **good** *(bad) news if $P(t = c|I) > (<)\pi$.*

I first justify the strong off-path belief.

Claim 1. *The divinity refinement in Banks and Sobel [1987] assigns the strong off-path beliefs in Definition 1.*

Proof. Following the notations in Fudenberg and Tirole [1991] I let $D((t,s),\mathcal{M},x')$ (and $D^0((t,s),\mathcal{M},x')$) be the set of the principal's mixed-strategy best response – the set of principal's retention probability – to the agent's action x' and belief concentrated on $\mathcal{M} \subset \{c,n\} \times \{g,b\}$ that makes a type (t,s) agent strictly benefits (and indifferent) by taking x' relative to his equilibrium action. Here $t \in T = \{c,n\}$ is the agent's payoff type; $s \in \{g,b\}$ is the agent's signal. The divinity condition says that fixing some off path action x', if for some $s \in \{g,b\}$ and $t \in \{c,n\}$ we have $D((t,s),\mathcal{M},x') \cup D^0((t,s),\mathcal{M},x') \subset \bigcup_{(t',s')\neq(t,x)} D((t',s'),\mathcal{M},x')$ then we can assign probability 0 that the deviation comes from type (t,s). We iterate this process if necessary until the divine equilibrium is found.

It is useful to note that (n,b) and (n,g) share the same preferences. Hence from now on I use $(n,\cdot)$ to denote these two types if they shall be preserved or ruled out together. Note also that we may arrange types $(c,g),(c,b),(n,\cdot)$ in the descending order of their motivation in reform. To establish our claim, we would like to 1) strike type $(n,\cdot)$ if the principal observes (r,e) for some $e \in [0,1]$, and 2) assigns probability 1 to the type $(n,\cdot)$ if the principal observes q. To save notations, I simply notate $x' = r$ if reform is off path and the agent in this deviation reforms with some effort $e \geq 0$.

Let $\underline{p}^{(c,s)}(\mathcal{M},x')$ be the retention probability that a congruent agent with signal s finds indifferent to his equilibrium payoff after he deviates to action x'; similarly we define $\underline{p}^{n}(\mathcal{M},x')$.

Then $D((t,s),\mathcal{M},x') = \{p : p > \underline{p}^{(c,s)}(\mathcal{M},x')\}$ and $D((n,\cdot),\mathcal{M},x') = \{\underline{p}^{n}((t,s),\mathcal{M},x')\}$ where p is the retention probability. To prove the claim, it suffices to verify whether $\underline{p}^{(c,s)}(\mathcal{M},q) > \underline{p}^{n}(\mathcal{M},q)$ for all s and $\underline{p}^{(c,s)}(\mathcal{M},r) < \underline{p}^{n}(\mathcal{M},r)$ for all s.

Consider the case in which q is off-path. This suggests that in equilibrium, every type of the agent reforms and the principal retains on path. Let $e^*_{(n,\cdot)}$ and $e^*_{(c,s)}$ be the equilibrium effort level of a noncongruent and a congruent agent with a signal s.

1) If efforts are observable then $e^* := e^*_{(n,\cdot)} = e^*_{(c,g)} = e^*_{(c,b)}$. To see this, if on path there are two different actions $(r,e),(r,e')$ with $e' > e$ such that the principal retains on both actions, then the noncongruent agent would always play one with the lower action (r,e). Then the action (r,e) becomes bad news about congruence and the principal should replace on path. Hence, any pooling equilibrium must take the form that both types pool with at unique action of the form (r,e).

2) If efforts are not observable then $e^*_n = 0$ and $e^*_{(c,s)} \in \arg\max \mu(s)e - \frac{e^2}{2\lambda}$ with $\mu(s) = \mu_+$ if $s = g$ and μ_- if $s = b$. This follows from the requirement of sequential rationality in any PBE.

Now, the definition of $\underline{p}$ requires that if efforts are observable, $\underline{p}^{(c,s)}(\mathcal{M},q)\cdot R + d = R + \mu e^* - \frac{e^{*2}}{2\lambda}$ for all s and $\underline{p}^{n}(\mathcal{M},q)\cdot R + d = R - \frac{e^{*2}}{2\lambda}$. If efforts are unobservable then $\underline{p}^{(c,s)}(\mathcal{M},q)\cdot R + d = R + \max_e\{\mu(s)e - \frac{e^2}{2\lambda}\}$ and $\underline{p}^{n}(\mathcal{M},q)\cdot R + d = R$. One concludes that in both cases $\underline{p}^{(c,s)}(\mathcal{M},q) > \underline{p}^{n}(\mathcal{M},q)$. In other words, the noncongruent has more to gain by deviating to the status quo policy than the congruent one during both good and bad times. This implies that we can strike (c,b) and (c,g) in sequence with the universal divinity refinement. The off-path belief about the payoff type t following $\{x = q\}$ surviving the divinity condition is noncongruent for sure.

The same logic applies to the case in which r is off-path – the congruent agent has more to gain from a deviation. Let's fix a pooling equilibrium at $x = q$ and consider the deviation to $x = r$ with any nonnegative effort e'. Since for any e' the type $(n,\cdot)$ always obtains strictly

less reform payoff than both congruent types (c, g) (c, b), whenever the deviation benefits some[10] congruent types the divinity condition strikes type $(n, \cdot)$. □

Now I describe the agent's behavior subject to different retention incentives.

Lemma 2. *Absent retention incentives, a noncongruent agent always take the status quo policy; a congruent agent initiates the reform with effort $\lambda\mu_+$ after $s = g$ and keeps the status quo after $s = b$.*

Proof. The noncongruent agent's optimal behavior is obvious. Let μ be the congruent type's posterior belief that the reform is good. Conditional on initiating a reform, his objective is

$$\max_e \mu e - \frac{e^2}{2\lambda}$$

The optimal effort is $\lambda\mu_+$ after $s = g$, and $\lambda\mu_-$ after $s = g$; The agent's reform payoffs are respectively $\frac{\lambda}{2}\mu_+^2$ and $\frac{\lambda}{2}\mu_-^2$. By Assumption 1, he initiates reform if $s = g$ and keep the status quo if $s = b$. □

Lemma 3. *Suppose the principal retains if and only if observing $(r, 1)$. Let μ be the agent's posterior belief about the state being $\omega = G$. Then conditional on a reform, the congruent agent exerts effort $\lambda(1 + R)\mu$ while the noncongruent agent exerts effort $\lambda R\mu$.*

Proof. Rewrite the agent's objective function as

$$\begin{aligned} &\text{Congruent} && \max_e \mu e(1 + R) - \frac{e^2}{2\lambda} \\ &\text{Nonongruent} && \max_e \mu eR - \frac{e^2}{2\lambda} \end{aligned}$$

The result follows immediately. □

10. For a deviation with prohibitively high efforts, the divinity condition has no bite. When this is the case, we may nonetheless assign the "strong off-path belief".

Lemma 4. *Assumption 2 and* $\lambda(1+R) \leq 1$ *imply* $R > 2d$.

Proof. Suppose the noncongruent agent secures retention conditional on a successful reform. His objective is $\max_{e\in\{0,1\}} \mu eR - \frac{e^2}{2\lambda}$ with μ being the belief that the reform is good. This means that the noncongruent agent's maximum payoff is $\frac{\lambda R^2\mu^2}{2}$. By Assumption 2, this value is larger than d if $\mu = \mu_+$. On the other hand, $\frac{\lambda R^2\mu^2}{2} \leq \frac{\lambda R^2}{2} < \frac{R}{2}$ since $\lambda R < \lambda(1+R) \leq 1$. This means that $R > 2d$. □

1.6.2 Equilibrium characterization

The proof of Fact follows directly from Lemma 2. Let us prove Results 1-3.

Observable policy

Proof. First check that the equilibrium described in Result 1 is indeed an equilibrium. With the strong off-path belief, the principal replaces whenever observing $x = q$. Since $R > d$, no agent would deviate from $x = r$. The agent chooses effort $\lambda\mu$ for each posterior belief $\mu \in \{\mu_-, \mu_+\}$ and the noncongruent agent chooses zero effort.

Second, I rule out other pure-strategy equilibrium possibilities under the strong off-path beliefs. 1) It cannot be the case that in equilibrium, one type of agent keeps the status quo more often than the others. Suppose in equilibrium the congruent type does q more often than the noncongruent type. Then $\{x = q\}$ is good news for retention and the noncongruent type would deviate to keeping the status quo. Suppose instead the noncongruent agent does q more often. Then $\{x = q\}$ is bad news and $\{x = r\}$ is good news for retention. Since $R > 2d$, the noncongruent agent would deviate to choosing r for all signals. 2) It cannot be the case that both types of agent takes $x = q$ regardless of signals: the congruent agent would initiate the reform with effort $\lambda\mu_+$ after a good signal and convince the principal that he is congruent and thus get reelected. 3) It cannot be the case that both types of agent chooses r after $s = g$ and chooses q after $s = b$. When this is the case, the noncongruent type would

deviate to choosing q after a good signal. Since $x = q$ is neutral news, the noncongruent type will be retained while enjoying his preferred policy. □

Observable outcome

First, I rule out cases in which the agent's policy choices are signal-invariant.

Per earlier discussions, it cannot be part of an equilibrium that both types choose $x = q$ regardless of signals. It cannot be part of an equilibrium that the congruent agent always chooses $x = r$ with some nonnegative effort and the noncongruent agent always chooses $x = q$; otherwise, the noncongruent type would deviate to $x = r$ to secure reelection.

Note also that it cannot be the case that both types choose $x = r$ and exert some signal-dependent efforts. When this is the case, the principal infers that $(r, 1)$ is good news and $(r, 0)$ is bad news about congruence. But a noncongruent agent wants to deviate from this strategy: after a bad signal, by choosing $x = r$ with effort $\lambda R\mu_-$ he obtains $\frac{\lambda}{2}R\mu_-^2$. By Assumption 2, this payoff is lower than the status quo payoff d.

Next, I consider cases in which the agent's policy choice reponds to signals.

Claim 2. *The following strategies cannot be part of an equilibrium: both types choose $x = q$ after $s = b$; they choose $x = r$ and exert nonnegative efforts after $s = g$.*

Proof. Following this strategy $\{x = q\}$ is neutral news about congruence. Hence a noncongruent type would deviate to $x = q$ after observing $s = g$. □

Claim 3. *The following strategy cannot be part of an equilibrium: the noncongruent agent always chooses $x = q$ and the congruent agent chooses $x = q$ after $s = b$ and chooses r with some nonnegative effort $e \geq 0$ after $s = g$.*

Proof. According to this strategy, $\{x = q\}$ is bad news about congruence. There are two cases to consider. First, the principal retains on policy. Then both types of agents would deviate to choosing $x = r$ and get retained. Second, the principal retains on outcome i.e.

whenever $y = (r, 1)$. When this is the case, the congruent agent wants to deviate to $x = r$ after $s = b$: in doing so, she obtains a payoff of $\frac{\lambda}{2}(1 + R)\mu_-^2$. By Assumption 2, this payoff is better than d. □

It is also straightforward to rule out pathological strategies in which the agent reforms when the signal is bad and takes the status quo when the signal is good. The only sensible strategy profile that may constitute an equilibrium is what is specified in Result 2.

Claim 4. *Under the strategy specified in Result 2,* $(r, 1)$ *is good news and* $(r, 0)$ *is bad news whenever the informativeness condition holds.*

Proof. $(r, 1)$ being good news follows from the fact that the congruent type always exerts more effort than the noncongruent type after a good signal. It remains to check when $(r, 0)$ is bad news.

By the Bayes' rule,

$$P(t = c|(r, 0)) = \frac{P(t = c, (r, 0))}{P(r, 0)}$$

and $P(t = c, (r, 0)) = P(t = c, s = g, 0) + P(t = c, s = b, 0)$

$$\begin{aligned}
P(t = c, s = g, 0) &= P(s = g, \omega = g)[1 - \lambda(1 + R)]\mu_+) + P(s = g, \omega = b) \\
&= \phi p[1 - \lambda(1 + R)\mu_+] + (1 - \phi)(1 - p) \\
P(t = c, s = b, 0) &= P(s = b, \omega = g)[1 - \lambda R\mu_+] + P(s = b, \omega = b) \\
&= \phi(1 - p)[1 - \lambda R\mu_+] + (1 - \phi)p \\
P(t = c, r, 0) &= 1 - \phi\lambda(1 + R)[p\mu_+ + (1 - p)\mu_-]
\end{aligned}$$

Likewise,

$$\begin{aligned} P(t=n,r,F) &= P(t=n,s=g,0) \\ &= P(s=g,\omega=g)(1-\lambda R\mu_+) + P(s=g,\omega=b) \\ &= \phi p(1-\lambda R\mu_+) + (1-\phi)(1-p) \end{aligned}$$

Since $P(t=c|r,0) \leq \pi \Leftrightarrow P(t=c,r,0) \leq P(t=n,r,0)$, we can rewrite the necessary and sufficient condition to

$$\begin{aligned} 1-\phi\lambda(1+R)[p\mu_+ + (1-p)\mu_-] &\leq \phi p(1-\lambda R\mu_+) + (1-\phi)(1-p) \\ \Leftrightarrow \qquad \phi(1-p) + p(1-\phi) &\leq \phi\lambda p\mu_+ + \phi(1-p)\lambda\mu_-(1+R) \\ \Leftrightarrow \qquad p[1-\phi-\lambda\phi\mu_+] &\leq \phi(1-p)[\lambda\mu_-(1+R)-1] \end{aligned}$$

Substitute in $\gamma = \frac{1-p}{p}.z = \frac{1-\phi}{\phi}$, the last inequality is $z - \lambda\frac{1}{1+\gamma z} \leq \gamma[\lambda(1+R)\frac{\gamma}{\gamma+z} - 1]$ as desired. □

Remark 3 (Technical)**.** *The informativeness condition is statistical. Note that the RHS is negative because* $\lambda(1+R) \leq 1$*. This means that for any* λ*,* $\exists\bar{z}$ *such that for all* $z \leq \bar{z}$*, we can find sufficient small* γ *such that the inequality holds. Put differently, we have one degree of freedom to choose an element in* (λ, R) *satisfying* $\lambda(1+R) \leq 1$ *such that the set of parameters supporting the informativeness condition is nonempty.*

The following lemma makes this point precise.

Lemma 5. *Suppose* $z < \lambda$*. Then* $\exists\bar{p} \in (\frac{1}{2}, 1]$ *that is independent from other parameters such that for all* $p \geq \bar{p}$*, the informativeness condition holds.*

Proof. Rearrange this inequality to $z \leq \lambda\frac{1}{1+\gamma z} + \gamma[\lambda(1+R)\frac{\gamma}{\gamma+z} - 1]$. Define $F(\gamma) = \lambda\frac{1}{1+\gamma z} + \gamma[\lambda(1+R)\frac{\gamma}{\gamma+z} - 1]$. Under the assumption $\lambda(1+R) \leq 1$, we claim that F is

decreasing in γ. To see it

$$\begin{aligned} F'(\gamma) &= \lambda[-\frac{z}{(1+\gamma z)^2} + (1+R)(1-\frac{z^2}{(\gamma+z)^2})] - 1 \\ &\leq \lambda(1+R)(1-\frac{z^2}{(\gamma+z)^2})] - 1 \\ &\leq (1-\frac{z^2}{(\gamma+z)^2})] - 1 < 0 \end{aligned}$$

This means that as long as $F(0) > z$, there must be some $\bar{\gamma} \in (0, 1]$ such that $F(\gamma) \geq z$ for all $\gamma \leq \bar{\gamma}$. The lemma follows by substituting in $\gamma = \frac{1-p}{p}$. □

Proof of Lemma 1. The results for λ, R and ϕ are immediate from inspecting the informativeness condition; the result for p follows from Lemma 5. □

Finally, let's verify that the strategies and beliefs specified in Result 2 indeed constitute an equilibrium.

Proof of Result 2. According to the strategies in Result 2, $(q, \emptyset)$ is bad news. Now that $(r, 1)$ is good news and $(r, 0)$ is bad news for congruence, the principal retains only after observing a successful reform. Now apply Lemma 3. □

Full transparency

Proof of Result 3. Let us verify that the strategies and beliefs in Result 3 indeed constitute a PBE. First, the noncongruent agent does not want to mimic a congruent one even when the signal is good. To see it, the noncongruent agent values retention at R and the status quo payoff at d; he is unwilling to initiate a reform if it entails a cost higher than $R - d$, which amounts to an effort $\sqrt{2\lambda(R-d)}$. This means that as long as the congruent agent is willing to exert an effort above this level, separation happens. Notate μ as a congruent agent's posterior belief that the state is good. His policy payoff is single-peaked at the $\lambda\mu$. This means that the congruent agent exerts effort $\max\{\sqrt{2\lambda(R-d)}, \lambda\mu\}$ in equilibrium.

Now we verify that this equilibrium survives the universal divinity refinement. Let's consider whether a deviation (r, e') with $e' \notin \{e_H, e_L\}$ may benefit anyone. Recall that there are three payoff types $(c, g), (c, b), (n, \cdot)$ arranged in the descending order with respect to the incentive to reform. Clearly, no profitable deviation may occur with $e' \geq e_H$. So we consider cases $e' \in (e_L, e_H)$ and $e' < e_L$.

- $e' \in (e_H, e_L)$. Then among the reformers, only the type (c, g) may benefit from this deviation when $e_H = \sqrt{2\lambda(R-d)} > \lambda\mu_+$ because it allows him to save effort for separation; the type (c, b) cannot benefit from this deviation. But the principal then regards this as bad news for retention. To see it, we again use the notation $\underline{p}^{(c,g)}$ and $\underline{p}^n$ to represent the retention probability that respectively make a type (c, g) agent and a type $(n, \cdot)$ agent indifferent between deviation and obtaining the equilibrium payoff; it suffices to show that $\underline{p}^n < \underline{p}^{(c,g)}$ i.e. the noncongruent agent benefits more from this deviation and thus can tolerate more retention loss. By definition, $\underline{p}^n \cdot R - \frac{e'^2}{2\lambda} = d$ and $\underline{p}^{(c,g)} \cdot R + \mu_+ e' - \frac{e'^2}{2\lambda} = R + \mu_+ e_H - \frac{{e_H}^2}{2\lambda} = d + \mu_+ e_H$. Since $e_H > e'$, straightforward comparison suggests that $\underline{p}^n < \underline{p}^{(c,g)}$. It happens because the noncongruent agent has less stake in reform and so he does not suffer as much as the congruent type in reducing efforts. As such, the principal must replace this agent (c, g) and the agent cannot benefit from this deviation.

- $e' \in (0, e_L)$. Then this deviation may benefit the (c, b) if $\lambda\mu_+ > e_L = \sqrt{2\lambda(R-d)} > \lambda\mu_-$ and the principal retains; it may benefit both types of (c, g) and (c, b) if $e_L = \sqrt{2\lambda(R-d)} > \lambda\mu_+$ the principal retains. We consider the first case; the second one follows analogously. The main observation goes just as above: whenever the congruent agent benefits from this deviation, the noncongruent type benefits more because he does not suffer from the loss of reform benefit. More precisely, $\underline{p}^n \cdot R - \frac{e'^2}{2\lambda} = d$ and $\underline{p}^{(c,b)} \cdot R + \mu(b) e' - \frac{e'^2}{2\lambda} = R + \mu(b) e_L - \frac{{e_L}^2}{2\lambda} = d + \mu(b) e_L$. Straightforward comparison shows that $\underline{p}^n < \underline{p}^{(c,b)}$. This suggests that the principal strikes (c, b) if she observes

any deviation to (r, e') with $e' \in (0, e_L)$ and believes that this deviation comes from type $(n, \cdot)$.

Hence, we have established the proposition. □

Lemma 6. *1. Among all separating equilibria, the universal divinity refinement selects uniquely on the least costly separating equilibrium described in Result 3.*

2. If $\lambda\mu_+^2 < 2(R-d)$, the universal divinity refinement cannot rule out a class of pooling equilibrium with the following properties: both types of agent choose $x = r$ with effort e^*, where $e^* \in [\lambda\mu_+, \sqrt{2\lambda(R-d)}]$. Otherwise, no pooling equilibrium may survive the universal divinity criterion.

Proof. **Part 1.** Result 3 describes the least costly separating equilibrium or the Riley outcome that survives the universal divinity refinement.

Consider other possibility of separation. First consider perfect separation in which an agent reforms if and only if he is congruent. We claim that the only plausible equilibrium in this category is the one characterized in Result 3. To see it, $\sqrt{2\lambda(R-d)}$ is the minimal effort to deter a noncongruent agent from mimicking. Any other equilibrium must involve the congruent exerting an effort weakly larger than this. However, any other separating equilibrium in which a type (c, g) chooses (r, e'_H) and a type (c, b) chooses (r, e'_L) with either $e'_H \neq e_H$ and/or $e'_L \neq e_L$ does not survive a deviation to the strategy specified in Result 3. For a concrete example, suppose towards contradiction that indeed a type (c, g) agent chooses (r, e'_H). Then he benefits most by deviating to (r, e_H); upon observing this observation, the principal believes according to the divinity condition that the agent has type (c, g). Consequently, the only equilibrium possibility is one described in Result 3.

Next we rule out "semi-separating" possibilities in which either 1) not all congruent types reform and all noncongruent types stay with the status quo or 2) not all noncongruent type stays with the status quo and all congruent types reform. In the first case, not reforming is

bad news about congruence. If on path the type (c, b) agent does not reform, the he may profitably deviate by reforming with effort $\lambda\mu_-$; upon this deviation the principal applies the divinity criterion and assigns probability 1 to type (c, b) and retains. Same story applies to type (c, g). In the second case, since the types of $(n, g), (n, b)$ share the same policy preference, they should behave the same in equilibrium. So we can rule out this possibility as well.

All other semi-separating equilibrium possibilities involving one of $(c, g)\&(c, b)$ reforms and one of $(n, g)\&(n, b)$ keeps the status quo can be easily ruled out.

Part 2. There are a few steps:

Step 1. Claim: In any pooling equilibrium that survives the divinity refinement, it must be that the agent reforms with an effort $e \geq \lambda\mu_+$.

Proof. Suppose not. This boils down to two possibilities: all types of agents $(c, g), (c, b), (n, \cdot)$ 1) pool on the status quo, or 2) pool on the reform with effort $e < \lambda\mu_+$. In either case, however, a (c, g) type can profitably deviate by choosing $(r, \lambda\mu_+)$. Upon this deviation, the divine condition assigns the probability 1 that this deviation comes from a type-(c, g) agent since he benefits more than other types. The agent will be retained, thus contradicting the equilibrium condition. □

Step 2: Fix a pooling equilibrium in which everyone reforms with effort $e^* \geq \lambda\mu_+$. Let's use the divinity condition to pin down the off-path beliefs. On path, every type shall be retained. For any deviation to be profitable it must be that either (1) the deviation involves the agent reforms with an effort $e' < e^*$, or (2) the agent chooses the status quo.

Case (1). As before define $\underline{p}^{(\cdot)}$ as the break-even retention probability after deviation.

By definition,

$$\begin{aligned} R\underline{p}^{(c,g)} + \mu_+ e' - \frac{(e')^2}{2\lambda} &= R + \mu_+ e^* - \frac{(e^*)^2}{2\lambda} \\ R\underline{p}^{(c,b)} + \mu_- e' - \frac{(e')^2}{2\lambda} &= R + \mu_- e^* - \frac{(e^*)^2}{2\lambda} \\ R\underline{p}^{n} - \frac{(e')^2}{2\lambda} &= R - \frac{(e^*)^2}{2\lambda} \end{aligned}$$

By the assumption that $e^* > e'$, we deduce $\underline{p}^{(c,g)} > \underline{p}^{(c,b)} > \underline{p}^{n}$; in other words, the noncongruent type benefits from the deviation the most. The the principal assigns probability 1 that the agent is of type-n following any deviation (r, e') with $e' < e^*$.

Case (2). Repeat steps in Case (1) and modify the deviation to q. By definition,

$$\begin{aligned} R\underline{p}^{(c,g)} + d &= R + \mu_+ e^* - \frac{(e^*)^2}{2\lambda} \\ R\underline{p}^{(c,b)} + d &= R + \mu_- e^* - \frac{(e^*)^2}{2\lambda} \\ R\underline{p}^{n} + d &= R - \frac{(e^*)^2}{2\lambda} \end{aligned}$$

As before, $\underline{p}^{(c,g)} > \underline{p}^{(c,b)} > \underline{p}^{n}$. The the principal assigns probability 1 that the agent is of type-n following any deviation (r, e') with $e' < e^*$.

Taken together, I have shown that any deviation that might benefit the agent would make the principal more suspicious that he is a noncongruent type.

Reversing the argument, it is straightforward to verify that the divinity condition assigns any unprofitable deviation to (r, e') with $e' \in (e^*, \sqrt{2\lambda(R-d)}]$ a belief that the agent is of type (c, g) with probability 1. Hence if $\lambda\mu_+ > 2(R-d)$, the following pooling equilibrium survives the divinity refinement: *All types of agent pool on the action* (r, e^*) *with* $e^* \in [\lambda\mu_+, \sqrt{2\lambda(R-d)}]$*; the principal assigns probability* 1 *that the agent is noncongruent upon observing* $x = q$ *or* (r, e') *with* $e' < e^*$*; and she assigns probability* 1 *that the agent is*

congruent & has received the signal $s = g$ upon observing (r, e') with $e' > e^*$. □

1.6.3 Welfare Comparison

Proof of Proposition 1

Proof. Across three information regimes, the congruent agent always initiates reforms. He exerts the least effort under the observable policy regime. To see why the congruent agent shirks most there, after taking the "correct" position he no longer worries about office. In other two regimes, the congruent agent has to either gamble for success or separate from the noncongruent type. Together with the fact that the noncongruent agent always fails a reform after exerting zero effort, the principal's policy payoff is the lowest under the observable policy regime.

To see why the observable outcome regime may prevail, it suffices to check whether the congruent agent works hardest under this regime. Were this true, then the principal would prefer the observable outcome regime when there is a sizable proportion of congruent agents in the pool (π high). A sufficient condition is $\lambda(1+R)\mu_- \geq \sqrt{2\lambda(R-d)}$ or equivalently $\lambda(1+R)^2\mu_-^2 \geq 2(R-d)$.

RHS is bounded above by $2(R-\frac{\lambda}{2}R^2\mu_-^2)$ using the condition $d \geq \frac{\lambda R^2\mu_-^2}{2}$ so it is sufficient to show that $R \leq \frac{\lambda}{2}\mu_-^2[R^2+(1+R)^2]$. This condition can be further simplified to $2 \leq \lambda\mu_-^2[2R+2+\frac{1}{R}]$. For sufficiently small R, we can always find $\lambda \in [0,1]$ satisfying this inequality and the parameter restriction $\lambda(1+R) \leq 1$. Further, per Remark 3 we can identify a set of parameters satisfying the informativeness condition. Finally, there exists a set of sufficiently small d satisfying Assumption 2.

There also exists parameters such that the full transparency regime prevails. Examples are available in the proof of Proposition 2. □

1.6.4 Comparative Statics

Proof of Proposition 2. **Part 1-2** follow from the reasoning in the text. The only caveat is that we need to verify whether $2(R-d) < \lambda$ *might be* consistent with Assumption 1-2, the informativeness condition, and $\lambda(1+R) < 1$.

I claim that there exist parameter values that are compatible with $2(R-d) < \lambda$ and the restrictions listed above. To see it, suppose that the signal accuracy is very high or $p \approx 1$; it overwhelms a weaker prior ϕ (e.g. $\phi = \frac{3}{4}$), resulting in $\mu_+ \approx 1$ and $\mu_- \approx 0$. Now I check these restrictions one by one.

1. With $p \approx 1$ Lemma 5 guarantees the informativeness condition if $z < \lambda$.
2. With $\mu_- \approx 0$, Assumption 1-2 reduce to $\max\{(1+R)\mu_-, R\mu_+\} > \sqrt{\frac{2d}{\lambda}}$ and $\mu_+ > \sqrt{\frac{2d}{\lambda}}$. If we pick $\lambda > \max\{\frac{2d}{R^2\mu_+^2}, \frac{2d}{\mu_+^2}\} \approx \max\{\frac{2d}{R^2}, 2d\}$ then these two assumptions hold.
3. We also need $\lambda(1+R) \leq 1$.
4. Need to verify that $\lambda\mu_+^2 > 2(R-d)$

Taken together, we may choose parameters like this: pick $p = \frac{99}{100}, \phi = \frac{3}{4}(z = \frac{1}{3}), \lambda = \frac{1}{2}, R = \frac{1}{4}, d = \frac{1}{80}$. This gives $\mu_+ \approx 0.996$ and $\mu_- \approx 0.03$ and $\sqrt{\frac{2d}{\lambda}} \approx 0.22$. $\lambda\mu_+^2 \approx 0.496$ and $2(R-d) = 0.475$. All restrictions are met.

Part 3. We want to construct a pair of vectors $w' = (p, \phi, d, \lambda, R, \pi), w'' = (p, \phi, d, \lambda, R', \pi)$ with $R' > R$ that satisfying Assumptions 1-2, the informativeness condition, and $\lambda(1+R) < 1$, such that the principal prefers the observable outcome regime under w'; she prefers the full transparency under w''.

To simplify matters, I assume that the agent is likely to be congruent ($\pi \approx 1$); I let $p = \phi = 1 - \epsilon$ for ϵ sufficiently small. Consequently, $\mu_+ \approx 1$ and $\mu_- = \frac{1}{2}$. With these two assumptions, the agent is unlikely to receive a bad signal (which occurs with probability $p(1-\phi) + \phi(1-p) \approx 2\epsilon$); the welfare comparison reduces to which information policy would elicit more efforts from a congruent agent when the signal is good.

Under the observable outcome regime, the congruent agent exerts effort $\lambda(1+R)\mu_+$. Under full transparency, he exerts $\max\{\sqrt{2\lambda(R-d)}, \lambda\mu_+\}$. The full transparency regime may elicit more effort whenever $\sqrt{2\lambda(R-d)} \geq \lambda(1+R)\mu_+$ or $\lambda\mu_+^2(1+R) \leq 2(R-d)$. Define $\hat{\lambda} = \lambda\mu_+^2$ and $H(R) = \hat{\lambda}(1+R)^2 - 2(R-d)$. H has real solutions if $1 \geq 2\hat{\lambda}(d+1)$; in this case, two solutions are $\underline{R} = \frac{1-\hat{\lambda}-\sqrt{1-2(1+d)\hat{\lambda}}}{\hat{\lambda}} > 0$ and $\bar{R} = \frac{1-\hat{\lambda}+\sqrt{1-2(1+d)\hat{\lambda}}}{\hat{\lambda}}$. Given the quadratic shape of H, for all $R < \underline{R}$ the observable outcome regime dominates; for $R \in (\underline{R}, \bar{R})$ full transparency dominates. This suggests that if R is close enough to $\underline{R}$ or $\bar{R}$, then the principal is willing to switch information regimes when there is small perturbation to R; if $R < \underline{R}$ or $R > \bar{R}$ then a local increase in R would not induce regime shift.

We claim from the expression $\underline{R}, \bar{R}$ that:

Claim 5. *$\underline{R}$ is increasing in λ and d; $\bar{R}$ is decreasing in λ and d.*

Proof. ($\underline{R}$): Since $\hat{\lambda} = \lambda\mu_+$ it suffices to verify $\frac{\partial \underline{R}}{\partial \hat{\lambda}} \geq 0$ and $\frac{\partial \underline{R}}{\partial d} \geq 0$. The latter is obvious. Note also that

$$\frac{\partial \underline{R}}{\partial \lambda} = \frac{\frac{\lambda(1+d)}{\sqrt{1-2(1+d)\lambda}} - (1 - \sqrt{1-2\lambda(1+d)})}{\lambda^2}$$

Denote $k = \sqrt{1-2(1+d)\lambda} \Leftrightarrow (1+d)\lambda = \frac{1-k^2}{2}$. The numerator of the above expression is $\frac{1-k^2}{2k} - (1-k) = \frac{1}{2}(k + \frac{1}{k}) - 1 \geq 0$.

($\bar{R}$): Similarly, $\frac{\partial \bar{R}}{\partial \lambda} \geq 0 \Leftrightarrow -\frac{\lambda(1+d)}{\sqrt{1-2(1+d)\lambda}} \geq \sqrt{1-2(1+d)\lambda}$ which is always false; the case for d is again straightforward. □

It remains to verify that the vector thus constructed $(p, \phi, d, \lambda, R) = (1-\epsilon, 1-\epsilon, d, \lambda, \bar{R})$ may satisfy all necessary assumptions. We have the freedom to choose d and λ. I let $\epsilon \downarrow 0$ for simplicity.

1. $\lambda(1+\underline{R}) \xrightarrow{\epsilon\downarrow 0} \lambda + 1 - \lambda - \sqrt{1-2(1+d)\lambda} < 1$.

2. The informativeness condition is guaranteed for $p = \phi = 1 - \epsilon$ and $\lambda > 0$. To see it, $z = \gamma = \frac{\epsilon}{1-\epsilon} \approx 0$ so the condition $z - \lambda \frac{1}{1+\gamma z} \leq \gamma[\lambda(1+R)\frac{\gamma}{\gamma+z} - 1]$ reduces to $\lambda > 0$.

3. Assumption 1 requires $1 > \sqrt{\frac{2d}{\lambda}} > \frac{1}{2}$ for ϵ arbitrarily small. This means that fixing λ it must be that $d \in (\frac{\lambda}{8}, \frac{\lambda}{2})$.

4. Assumption 2 simplifies to $\max\{\frac{1+\underline{R}}{2}, \bar{R}\} > \sqrt{\frac{2d}{\lambda}} > \frac{\underline{R}}{2}$. Unpacking the expression $\underline{R} = \frac{1-\hat{\lambda}-\sqrt{1-2(1+d)\hat{\lambda}}}{\hat{\lambda}}$, letting $\hat{\lambda} \to \lambda$ (since $\epsilon \downarrow 0$), a sufficient condition is

$$1 - \lambda - \sqrt{1 - 2(1+d)\lambda} < 2\sqrt{2d\lambda} < 1 - \sqrt{1 - 2(1+d)\lambda}$$

There is a large set of pairs (d, λ) satisfying the above three conditions. For example, we can let $d = 0.05$ and $\lambda = 0.3$. Condition 1, 2 and 3 are immediate. Condition 4 simplifies to $0.0917 < 0.346 < 0.39$.

Finally, from $w = (1-\epsilon, 1-\epsilon, 0.05, 0.3, \bar{R})$ we may construct $w' = (1-\epsilon, 1-\epsilon, 0.05, 0.3, \bar{R}-\delta)$ and $w'' = (1-\epsilon, 1-\epsilon, 0.05, 0.3, \bar{R}+\delta)$ with $\delta > 0$ sufficiently small such that the principal chooses the observable outcome regime under w' and full transparency under w''. □

1.6.5 Robustness

Selection

The baseline model makes a simplifying assumption that the principal is entirely policy-motivated. One may wonder to what extent this assumption drives my results; after all, it is quite reasonable to assume that the principal may attach nontrivial weight on selecting a congruent agent.

I argue that all results continue to hold as long as the principal's weight on selection is not too high. Crucially, note that the principal's retention happens at the last stage. Since the principal retains on good news about congruence, her retention strategy remains invariant

to the weight of selection. Also note that a policy-motivated principal's optimal information policy is generically[11] unique. This means that even if this principal starts to care about selection, she would not change her policy if selection is not that important.

Modeling congruence

I model (non)congruence along the line of Fox [2007]. One may wonder whether alternative notions of congruence (e.g. Maskin and Tirole [2004]) might induce similar or different results.

I argue that the current setup presents the motivation issue in a transparent way. By contrast, an alternative setup closer to Maskin and Tirole [2004] often involves the noncongruent type discounting future differently than the congruent type. The discriminatory nature of career concerns in the latter setup blurs the comparison of motivation and separation effects across different information regimes.

To see this, suppose the noncongruent type is a reform saboteur – he prefers a failed reform to the status quo to a successful reform. Per Maskin and Tirole [2004], we would like this type of agent to trigger a reform whenever the reform timing is bad, and stays with the status quo whenever the timing is good in the absence of career concerns. This equilibrium behavior could be induced, for example, when efforts and the policy choices are *substitutes* for reform success: a good reform always succeeds, and a bad reform succeeds with a probability equal to effort. The main issue is that, the agent cannot control the (risky) reform outcome in a deterministic way; this renders asymmetry in the continuation payoffs to two types of agent. In the modified setup above, the congruent type at least secures $R+\phi$ by holding office in the period 2 (reform without effort); the noncongruent type obtains at most $R + (1 - \phi)$. The congruent agent benefits more from holding office at least when the prior is biased towards reform ($\phi \geq \frac{1}{2}$). In this case, career concerns discipline the congruent

11. The set of parameters that give rise to more than one optimal policy is meager in the parameter space.

agent more strictly than the noncongruent one; accordingly, it makes type-separation easier relative to the main model. While this observation lends extra credibility that an opaque information regime may prevail (since the congruent agent exerts less effort on path under full transparency), it is not obvious whether its optimality comes mainly from the motivation effect of a pivotal decision rule or just the asymmetric discipline effect.

CHAPTER 2

GENERALIZED PEACEFUL MECHANISMS

2.1 Introduction

Mediation is traditionally regarded as an invaluable information channel for states to resolve conflicts. An oft-cited example is the shuttle diplomacy of the 1990 Kashmir crisis. In April 1990, Washington warned about "a growing risk of miscalculation which could lead events to spin out of control", and later dispatched Robert Gates to avert the Indo-Pakistani crisis [Hagerty, 1998, 150]. The key to the Gates' success, Hagerty argues, lies in his persuasive message to both sides that the war would be detrimental. He also notes that Gates promised to monitor and share information about states' compliance using American spy satellites. In doing so, the Gates mission reduced the chance of war that may arise from India and Pakistan's miscalculation of information uncertainty (as in Blainey [1988] and Fearon [1995]).

Gates' success notwithstanding, mediation is generally hard. Even the received wisdom that mediators facilitate information transmission is not unconditional. A prerequisite is that states must find the mediator trustworthy before they are willing to share private information [Kydd, 2006]. But trust alone is not enough. Fey and Ramsay [2010] point out that if a mediator does not strategically process the information, then states would again have incentives to misrepresent private information as if the mediator were absent. This suggests that a mediator must overcome states' truth-telling constraints in order to solve the information problem in mediation.

To incentivize peaceful settlements, mediators must also account for political leaders' domestic constraints. Political leaders represent their domestic constituencies in crisis bargaining, but they do not necessarily share constituencies' preference for the bargaining outcome. The preference divergence may come from leaders' disproportional gain and/or loss relative to their fellow citizens in wars, also known as the *political bias* [Bueno de Mesquita

et al., 1999, Jackson and Morelli, 2007]; it may also occur during peacetime taking the form of *audience costs*: after backing down from crises, political leaders incur a penalty by their domestic constituencies [Fearon, 1994, Schultz, 2001, Tarar and Leventoglu, 2013]. How domestic constraints interact with the prospect of mediation is less well understood by the literature. Ashworth and Ramsay [2017] analyze the optimal domestic constraint design in crisis bargaining situations, but leave it open whether peacefully resolving crises is possible.

Finally, mediators have to contemplate whether states would respect the designated peaceful settlement. This is not a concern if mediators have the power to enforce agreements[1]. In reality, however, "even powerful intermediaries rarely can impose a settlement; their mediation efforts are constrained by circumstances" [Kriesberg, 2001, 378]. There lacks a convincing counterfactual to the Indo-Pakistani crisis regarding how US would have intervened had either side reneged, but we do learn from the Arab-Israeli conflict conflicts. According to Kriesberg [2001], ten major events of conflict transformation occurred in the 1990s despite numerous mediation efforts.

This paper aims to reconcile the above three concerns, and examine the limits of mediated conflict resolution. I ask: when does a peaceful settlement exist? How do domestic constraints shape the scope of peace? And, how does a mediator's commitment power affect the implementation of a peaceful settlement?

My approach. I build on the crisis bargaining games developed by Banks [1990] and Fey and Ramsay [2011], and examine the availability of peaceful mechanisms with the game-free approach. The novel ingredient is that I introduce audience cost and political bias into the canonical crisis bargaining situations. Accordingly, I define a mediated mechanism to be peaceful if it involves a settlement such that for *some* audience cost, all stakeholders within each state would voluntarily accept it subject to the enforcement of a powerful mediator.

1. Strong mediators, even if they are biased, can design rules for interaction and enforce agreements with credible punishment [Goltsman et al., 2009, Hörner et al., 2015, Favretto, 2009]. Sometimes, mediators have the power to set agenda and influence the crisis bargaining outcomes [Camiña and Porteiro, 2009].

Here I fix political bias mainly for realistic concerns, because people do not always equally benefit or suffer from a conflict. I impose no restrictions on the functional form of audience costs. The purpose for doing so is to understand whether and how audience costs may impact peace possibilities, other than enabling leaders to learn their adversary's preference between war and peace [Fearon, 1994]. Taken together, this approach identifies the best scenario for states to resolve a crisis peacefully when domestic constraints are at play.

My game-free approach isolates the impact of information asymmetry from domestic constraints in studying the peace possibility. Its analytical power derives from the revelation principle, which allows us to analyze a class of outcome-equivalent bargaining procedures in which players have no incentives to misrepresent private information [Myerson, 1979, Myerson and Satterthwaite, 1983, Banks, 1990, Fey and Ramsay, 2011]. By examining whether these procedures may induce peaceful outcomes, I would characterize the peace condition. Indeed, the validity of this game-free approach rests on assuming a mediator who enforces any agreement in the background. From this perspective, my analysis establishes an ideal "strong mediator benchmark": if a mediator with enforcement power cannot implement some settlement template, then neither can mediators without enforcement power. Conversely, for settlements implementable by strong mediators, we may look for conditions under which they may also be implemented by weak mediators.

Main results. I characterize the peaceful mechanisms existence condition as a "weighted resource budget constraint": peace is possible whenever the total size of resource being divided exceeds the sum of two strongest-type states' expected war payoffs. This characterization technically enriches the peace constraint of Fey and Ramsay [2011] with domestic politics. Intuitively, it says that when the size of pie is large enough, a mediator may find some ways of redistribution to simultaneously pacify two states; otherwise, there must always be some strong types who cannot be satisfied via a peaceful settlement for *any* audience cost.

As an immediate consequence, we learn that political bias directly determines the peace

possibility. To see this, note that war is states' alternative option to resolve a crisis other than the mediated peace talk. Since political bias together with the war technology dictates how much each stakeholder may possibly obtain from war, it affects states' willingness to participate in the peace talk.

By contrast, audience costs do not discipline leaders' behaviors in *any* peaceful mechanism. The crucial observation is that, for leaders not to misrepresent private information in a peaceful mechanism, they must receive constant payoffs regardless of types. The observation exploits the property that audience costs are realized only upon peaceful settlements. As such, from Myerson's lemma [Myerson, 1981] we deduce that leaders' equilibrium payoff is completely pinned down by the type-dependent war probability ("allocation rule"), up to a constant determined by the strongest-type leader's payoff ("participation constraint"). Since peaceful mechanisms preclude the war possibility, leaders must receive the same payoff so long as they remain in the peace talk. Moreover, this payoff must exceed leaders' option value of war for the highest possible types. This means that in any peaceful mechanism, the minimum payoff that a leader would ask for is independent from audience costs. Put differently, audience costs are irrelevant for incentivizing peaceful settlements.

From characterizing the peace condition, I also relate leaders' private types to their war propensity and equilibrium payoff. The set of empirically-relevant monotonicity results says that, as a leader gets stronger, she/he tends to use military force more often, and obtains high equilibrium payoffs. The main intuition comes from leaders' incentive to misrepresent their types: because stronger leaders can always misrepresent and fight with the same probability as if they were weaker ones, being better fighters their equilibrium payoffs cannot decrease in types. Moreover, stronger leaders tend to exploit their military strength by demanding more concessions at the bargaining table at a higher risk of war. When war cannot be averted – which we know by the contrapositive of the peace condition – the monotonicity result must be strict over some range of types. Empirically, this means that we can deduce leaders'

private types by observing their equilibrium war-initiation behaviors.

To complete the game-free analysis, I describe how a mediator's enforcement power affects the implementation of peaceful settlements. With the power to enforce agreements, a strong mediator can designate any settlement within the bargaining range to resolve the crisis. By contrast, a weak mediator who lacks this power cannot prevent strong leaders from renegotiating for better terms of settlement at a higher risk of war. This suggests that weak mediators will find it much more difficult to preclude the war possibility. An exceptional case is that leaders at the bargaining table can credibly reject any renegotiated offer. The credibility may be supported by leaders' off-path belief – they expect any renegotiated offer to come from "weak enemies", and prefer to fight rather than accepting the terms of renegotiation. When this is the case, a weak leader faces a similar implementation environment as if she has the enforcement power.

I conclude the introduction section by relating my paper to existing works. This paper builds on and extends the mechanism design approach of crisis bargaining, and makes robust predictions for general crisis bargaining games. Banks [1990] first recognizes the necessity of an "institution-free" analysis for game-theoretic models, and establishes weak monotonicity theorems about leaders' war-propensity and payoff. Fey and Ramsay [2009, 2011] generalize Banks [1990]'s information structure, and study the possibility of peaceful resolution. Hörner et al. [2015] take yet another angle, suggesting that arbitration and mediation could yield the same resolution outcome with mediators' clever strategy of communication. This paper sharpens Banks [1990]'s weak monotonicity theorems, generalizes Fey and Ramsay [2009, 2011]'s peace condition, and studies peaceful mechanism implementation with respect to neutral mediators of different enforcement power in the spirit of Hörner et al. [2015].

The rest of the paper is organized as follows: Section 2.2 presents the structure of the model, which is analyzed in Section 2.3. Section 3.3 concludes.

2.2 Model

The model concerns two states disputing over a unit size of resource. Each state has a population normalized to 1 with share γ_i being leaders and share $1-\gamma_i$ being citizens. Leaders bargain in the shadow domestic constraints. At the outset, there is a mediator whose objective is to preclude any war possibility.

Mediated crisis bargaining

Leaders engage in a crisis bargaining game, whose final outcome is either a peaceful settlement or a war. The game entails the standard structure of information uncertainty about resolve[2], which is described by a set of types $\Theta = \Theta_1 \times \Theta_2$ with its typical element being $\theta = (\theta_1, \theta_2)$. θ_i and θ_j are independently distributed. Each state knows its own type $\theta_i \in \Theta_i = [\underline{\theta}_i, \bar{\theta}_i]$, but treats state-$j's$ type as a random variable with cumulative distribution function $F_j(\theta_j)$ and strictly positive density.

The crisis bargaining outcome is a tuple (π, x) specifying war probability $\pi \in [0,1]$ and peaceful division $x = (x_1, x_2) \in \Delta^2 = \{(x_1, x_2) : x_1 + x_2 \leq 1, x_1, x_2 \in \mathbb{R}_+\}$. War is a costly lottery with each side endowed with winning probability $p = (p_1, p_2)$ which sums to one. Leaders also suffer from war costs $c = (c_1, c_2)$ adjusted by their political biases. The winner takes all the resources.

θ determines state-i leader's war payoff $w_i(\theta)$, which is strictly increasing and differentiable in its argument θ_i. According to Fey and Ramsay [2011], the resolve θ_i may concern (one-sided) fighting costs $w_i(\theta) = w_i(-c_i)$ where $c_i \sim F_i(c_i)$ and w_i is increasing in $-c_i$; or it may concern (two-sided) relative strength in fight $w_i(p_i(\theta)) = w_i(p_i(\theta_i, \theta_j))$ where p_i is increasing in θ_i but decreasing in θ_j, and w_i is increasing in p_i.

Now denote leaders' strategy set abstractly as $S = S_1 \times S_2$. S itself is enormously large, but it contains a small subset $\Theta = \Theta_1 \times \Theta_2$ which coincides with leaders' type space.

2. See Ramsay [2017] for a review.

The class of game form known as the *direct mechanism* restricts each leader's strategy to reporting some type $\tilde{\theta}_i \in \Theta_i$, and it assigns war probability $\pi(\tilde{\theta})$ and peaceful settlement $x(\tilde{\theta}) = (x_1(\tilde{\theta}), x_2(\tilde{\theta}))$ for a typical report profile $\tilde{\theta} \in \Theta$. Formally, we represent the direct mechanism as a "menu" $\Sigma : \tilde{\theta} \to (\pi(\tilde{\theta}), x(\tilde{\theta}))$. It is incentive compatible (IC) if leaders of all types are willing to report truthfully (i.e. $\tilde{\theta}_i = \theta_i$).

It turns out that studying the class of incentive compatible direct mechanisms Σ would not compromise our search for the peace possibility. We may bypass the myriad of bargaining strategies thanks to the powerful revelation principle [Myerson, 1979]. Its IR version says that for any crisis bargaining game, we can construct an outcome-equivalent game form in the class of the incentive compatible direct mechanisms Σ [Fey and Ramsay, 2011]. Along with the fact that Σ is just a subset of all possible crisis bargaining game forms, studying its equilibrium outcome entails no loss of generality.

At this point, the common game-theoretic approach would be positing an exact game form so that we may analyze its equilibrium outcome. The following game form Γ is an example: first, the mediator offers and commits to enforcing a "menu" Σ to the leaders of two states-in-conflict. Second, each leader decides some $\tilde{\theta}_i \in \Theta_i$ to report. Third, the outcome $(\pi(\tilde{\theta}), x(\tilde{\theta}))$ realizes. If the menu satisfies the IC constraint, then we should expect $(\pi(\theta), x(\theta))$ to occur in equilibrium. If the goal is to examine when a mediator may mediate the crisis, we ask whether she/he can offer an *incentive compatible* menu inducing zero war possibility, namely $\pi(\theta) \equiv 0$ for all $\theta \in \Theta$.

While the game-free approach does not require positing any particular game form, we may use Γ as a concrete template to think about the mediation process. On the one hand, it suggests a conflict resolution procedure that a mediator may use whenever she/he knows that a peaceful settlement exists. On the other hand, the game form Γ does not place any additional restriction on the peace possibility thanks to the revelation principle. To see it, any particular game form must induce some final equilibrium outcome $(\pi(\theta), x(\theta))$. If a

mediator cuts through the complicated bargaining process and directly offers an *incentive compatible* menu $(\pi(\theta), x(\theta))$, then leaders cannot improve their payoffs by making unilateral deviations; otherwise, either the truth-telling or the equilibrium condition will be violated. In words, Γ is a valuable scaffold organizing key elements of the mediated crisis bargaining process, because it together with the IC constraint can replicate the equilibrium outcome of any game form.

Remark. The mediator's power to enforce the menu $(\pi(\theta), x(\theta))$ plays a key role here; absent the enforcement power, leaders would not consider the menu credible. While the enforcement power assumption might not be realistic (as argued in Kriesberg [2001]), it helps produce a set of peaceful settlements that is weakly larger than what the real-world implementation constraint would permit. We may then drop this assumption and verify whether these settlements are also implementable by weak mediators. In the section of implementation, I discuss this guess-and-verify approach in detail.

Domestic constraints

Audience cost

Formal literature [Fearon, 1994, Schultz, 2001, Tarar and Leventoglu, 2013, Ashworth and Ramsay, 2017] models audience costs as leaders' suffering if they initiate crises but end up backing down. Following this tradition, I model audience costs as leaders' "peace penalties", because accepting a peaceful settlement is inconsistent with leaders' initial intention of crisis escalation. After punishment, state i leader's ex post net gain would be $v_i(\theta)$ conditional on type $\theta \in \Theta$. The audience cost is naturally understood as the discrepancy between state $i's$ commonly-valued settlement and leader's net payoff $a_i(\theta) = x_i(\theta) - v_i(\theta)$, which encompasses all possible functional forms in the literature.

Citizens do not benefit directly from the audience cost a_i, because it is privately "valued" by the leader. Instead, citizens' peaceful payoffs are implicitly determined by the bargaining

outcome x_i. To summarize, with audience cost the leader and citizens' payoffs in state i are $v_i(\theta)$ and $x_i(\theta)$ conditional on type θ.

Political bias

Political bias is the war payoff discrepancy between leaders and citizens [Jackson and Morelli, 2007, Bueno de Mesquita et al., 1999, Ashworth and Ramsay, 2017]. If state-i citizens' war cost is c_i, then their leader suffers from $\lambda_i c_i$, where $\lambda_i > 0$ parametrizes the leader's cost sensitivity to conflict relative to citizens. To allow for the possibility of dovish leaders I do not necessarily restrict $\lambda_i \in (0, 1)$. Hence, if war occurs, a risk-neutral leader gets a net payoff $w_i = p_i - \lambda_i c_i$ while a citizen gets $w_i^c = p_i - c_i$.

Following Ashworth and Ramsay [2017] I take political bias as given, but treat domestic audience cost as a "free parameter" throughout the paper. This approach is substantively motivated by the fact that audience costs are more constitutional than political bias, as the latter is structurally shaped by the war technology beyond audiences' control. When a peaceful settlement exists for *some* audience cost, I consider peace as a plausible mediated bargaining outcome.

Participation constraints

Given the anarchic structure of the international system, states may back out from any signed peaceful agreement. They are willing to stay in the agreement only if doing so would bring a payoff higher than what they would expect from using force. The "voluntary participation constraints" reflect the concern of Waltz [2001], who suggests that "sovereign states with no system of law enforceable among them, with each state judging its grievances and ambitions according to the dictates of its own reason or desire".

In the IR context, the suitable participation constraint is the "interim individual rationality" defined in Banks [1990] and Fey and Ramsay [2011]. Its original formulation is a

statement about states' preferences between war and peace after knowing their own private types but not their opponents'. It requires that after making a type report (not necessarily truthful), each state receives a peaceful settlement weakly larger than its expected war payoff. In a peaceful mechanism, the participation constraint applies to each type of states.

After introducing domestic politics, I consider the participation constraints for the leader and citizens respectively. This approach reflects the view that each stakeholder has a say in signing a bilateral peaceful settlement. On the one hand, leaders may at any time walk away from the bargaining table and initiate a conflict. On the other hand, citizens might not entirely retire from their homeland's war-or-peace decision. For example, a leader has to garner enough support from soldiers before waging a war. Conversely, a peace treaty may not last long if it fails to pacify domestic citizens.

Now let us formalize the participation constraints. Denote stakeholders' the expected (interim) war/peace payoff as

$$\text{(War)} \qquad W_i(\theta_i) = \int_{\underline{\theta}_j}^{\bar{\theta}_j} w_i(\theta_i, \theta_j) dF_j(\theta_j), \quad W_i^c(\theta_i) = \int_{\underline{\theta}_j}^{\bar{\theta}_j} w_i^c(\theta_i, \theta_j) dF_j(\theta_j),$$

$$\text{(Peace)} \qquad V_i(\theta_i) = \int_{\underline{\theta}_j}^{\bar{\theta}_j} v_i(\theta_i, \theta_j) dF_j(\theta_j), \quad X_i(\theta_i) = \int_{\underline{\theta}_j}^{\bar{\theta}_j} x_i(\theta_i, \theta_j) dF_j(\theta_j),$$

Following Banks [1990], I define leaders and citizens' interim participation constraints as follows: for each $\theta_i \in [\underline{\theta}_i, \bar{\theta}_i]$

$$\text{(Leader)} \qquad W_i(\theta_i) \leq V_i(\theta_i) \tag{2.1}$$

$$\text{(Citizens)} \qquad W_i^c(\theta_i) \leq X_i(\theta_i) \tag{2.2}$$

The set of inequalities (2.1)-(2.2) describes the most stringent participation constraints in a bargaining environment with domestic politics. Substantively, it says that a peaceful agreement needs approvals from all domestic stakeholders before it comes into effect. While

the formal analysis of this paper devotes to this case, we can redefine participation constraints to accommodate situations in which not all stakeholders may veto a peaceful agreement. For example, if state-$i's$ pivotal decision maker is its political leader, then it suffices to consider this leader's participation constraint only.

Closing the model

In the remainder of this section, I formally describe players' payoff functions and discuss the method of the game-free approach. Each player in this game is risk-neutral. Let u_i, u_i^c stand for the *ex post utility* of the leader and a typical citizen in state i. These utilities are average payoff of war (w_i, w_i^c) and peace (v_i, x_i) weighted by "war assignment function" π, conditional on each type realization $\theta \in \Theta$:

$$u_i(\theta) = \pi(\theta)w_i(\theta) + (1 - \pi(\theta))v_i(\theta)$$
$$u_i^c(\theta) = \pi(\theta)w_i^c(\theta) + (1 - \pi(\theta))x_i(\theta)$$

Let $U_i(\tilde{\theta}_i|\theta_i)$ denote type-θ_i leader's expected (interim) utility if she reports $\tilde{\theta}_i$ in the mechanism,

$$U_i(\tilde{\theta}_i|\theta_i) = \int_{\underline{\theta}_j}^{\bar{\theta}_j} u_i(\tilde{\theta}_i, \theta_j)dF_j(\theta_j) = \int_{\underline{\theta}_j}^{\bar{\theta}_j} [\pi(\tilde{\theta}_i, \theta)w_i(\theta_i, \theta_j) + (1 - \pi(\tilde{\theta}_i, \theta_j))v_i(\tilde{\theta}_i, \theta_j)]dF_j(\theta_j)$$

and denote her utility from truthfully reporting as $U_i(\theta_i) := U_i(\theta_i|\theta_i)$.

So far, I have assembled all necessary elements for the game-free analysis. The main goal is checking whether a mediator with enforcement power may offer an incentive compatible menu $\Sigma : \theta \to (\pi(\theta), x(\theta))$, such that all domestic stakeholders would like to accept and resolve the crisis peacefully. To concretely think of the complete mediation process, we may augment the mediated crisis bargaining game Γ with domestic politics, and define an extensive-form game Ω. Ω begins with a mediator committing to offering a peaceful menu

Σ (i.e. $\pi \equiv 0$) to leaders at the bargaining table. If any leader rejects the menu, war ensues. Otherwise, each leader makes a type report $\hat{\theta}_i \in \Theta_i$ and brings the settlement $x_i(\hat{\theta}_i)$ to citizens for ratification; under the incentive compatible constraint, the equilibrium report must be $\hat{\theta}_i = \theta_i$. If citizens of both states ratify the settlement, then peace realizes; otherwise, war ensues. Domestic stakeholders' payoffs are given as before. The mediator wants to maximize the peace probability, and our goal is to examine whether this probability could be 1. Per earlier discussion, Ω works as well as the game-free approach in finding peaceful settlements.

It is worth mentioning that I take a "minimum" set of peaceful criteria for the mediated crisis bargaining situation. When a peaceful mechanism exists, it does not mean that peace is readily achievable; instead, it suggests that domestic stakeholders may institutionalize appropriate audience costs to resolve the crisis peacefully (perhaps with the help of a mediator). For example, citizens may either commit to a reward-punishment scheme [Barro, 1973, Ferejohn, 1986, Ashworth and Ramsay, 2017] or condition their retention strategy on an endogenous reference point [Acharya and Grillo, 2019] to discipline leaders' behaviors. But if none of mechanisms meets the minimal "peaceful" criterion, *any* peaceful settlement must be suboptimal for at least one of the states-in-conflict.

Finally, I comment on the role of the mediator in the mediation process. Because this paper aims to understand the limits of mediated conflict resolution, here the mediator cares only about whether she/he may make a menu Σ to preclude the war possibility. This approach differs from Kydd [2006], who assumes that mediators care about the issue at stake. As to whether the mediator remains "neutral" about the issue during mediation, my approach resembles Hörner et al. [2015], although their objective is to examine how a weak mediator may play clever communication strategies to circumvent the unenforceability constraint. I also rule out the possibility that the mediator may subsidize two states-in-conflict for peace. I do so mainly to avoid repetition: Fey and Ramsay [2011] show that

peace is always possible if a mediator is willing to make sufficiently large subsidies to appease both states. However, the size of subsidies could be too large to be politically feasible.

With these terminologies, it is possible to conduct the game-free analysis.

2.3 Analysis

Characterization

In this section, I derive the conditions for the existence of an incentive-compatible menu $\Sigma : \theta \to (\pi(\theta), x(\theta))$ that a mediator may offer to induce peace. The menu Σ is incentive-compatible if each leader at the bargaining table does not gain by misreporting her/his private information θ_i; it induces peace if 1) $\pi(\theta) \equiv 0$ for all $\theta = (\theta_1, \theta_2)$ up to a measure-zero set, and 2) whenever a mediator with enforcement power proposes it, all stakeholders always prefer to accept the peaceful agreement $x(\theta)$ specified in the menu rather than wage war.

The first result formalizes the incentive compatible condition of the menu Σ.

Lemma 7. *Σ is incentive compatible if and only if*

$$U_i(\theta_i) - U_i(\underline{\theta_i}) = \int_{\underline{\theta_i}}^{\theta_i} \int_{\underline{\theta_j}}^{\bar{\theta_j}} \pi(t, \theta_j) \frac{\partial w_i(t, \theta_j)}{\partial t} dF_j(\theta_j) dt, \tag{2.3}$$

$$\int_{\underline{\theta_j}}^{\bar{\theta_j}} \pi(t, \theta_j) \frac{\partial w_i(t, \theta_j)}{\partial t} dF_j(\theta_j) \text{ is weakly increasing in } t. \tag{2.4}$$

The proof follows the standard mechanism design technique and is deferred to the Appendix 2.5.1. Roughly, Condition 2.3 corresponds to leaders' first-order condition which precludes any local profitable deviation from truthful reports. Condition 2.4 establishes a kind of second-order condition that ensures the global optimality of truth-telling.

Conditions 2.3 and 2.4 imply the monotonicity of payoff and war propensity with respect to types. To see why, given the menu Σ the stronger leaders can always misrepresent as

weaker types; in doing so, they obtain the weaker one's assigned war probability and peace payoff. This means that the stronger leaders cannot be worse off than the weaker ones, because they are the better fighters whenever war happens. To induce truth telling, it must be that the stronger leaders obtain an even higher overall "rent" specified by Condition 2.3. At the same time, the weaker leaders must be deterred from misreporting upward. Condition 2.4 guarantees it, because by reporting upward the weaker leaders have to fight more often. As worse fighters, they are not as aggressive as the stronger types in trading off the risk and returns.

At first glance, the envelope formula seems to be a technical generalization of Banks [1990]. It says that in any mechanism, leaders' equilibrium payoff is completely pinned down by the war assignment function π and war technology w_i up to a constant. However, the envelope formula has surprising substantive implications: after introducing domestic constraints, the audience costs do not explicitly enter leaders' payoff function. *Prima facie*, the result seems to contradict Fearon [1994]'s notion of audience cost since it is payoff irrelevant. I remark that the audience costs implicitly impact leaders' war propensity: unless the war assignment function π does not depend on leaders' report, the standard risk-return trade off foreshadowed in Banks [1990] persists.

In the class of peace mechanisms, the IC constraint has a particularly simple structure. Because a typical peaceful menu prescribes the war probability $\pi(\theta) \equiv 0$ for all θ, every leader must receive the same payoff regardless of their true types. This means that $U_i(\theta_i) = U_i(\bar{\theta}_i)$ for all $\theta_i \in [\underline{\theta}_i, \bar{\theta}_i]$. The pooling payoff structure eliminates low types' incentive to mimic high types for bargaining gains *regardless of* leader-citizen preference divergence à la audience costs. To glorify the observation that leaders remain unresponsive to their true types in *any* peaceful settlement, I put it as one of the main results:

Result 4. *Audience costs are irrelevant for incentivizing peaceful settlements.*

With the help of Lemma 7 and the peace requirement that $\pi \equiv 0$, I rewrite the par-

ticipation constraints (2.1)-(1). At the bargaining table, leaders of all types must prefer accepting the peaceful settlement to simply fighting. Each leader understands that she/he can on average take $V_i(\theta)$ away by signing the treaty and $W_i(\theta_i)$ by fighting. In a peaceful mechanism, $V_i(\theta)$ coincides with one's expected payoff $U_i(\theta_i)$, and equals $U_i(\bar{\theta}_i)$ by the incentive compatibility requirement. As such, leaders' participation constraints satisfy the following chain of inequality: $\forall \theta_i \in [\underline{\theta_i}, \bar{\theta}_i]$,

$$\text{(Leader)} \qquad W_i(\theta_i) \leq V_i(\theta_i) = U_i(\theta_i) = U_i(\bar{\theta}_i)$$

Because the strongest leader receives the most spoils from the battlefield, it must be that $\sup_{\Theta_i} W_i(\theta_i) = W_i(\bar{\theta}_i)$. This suggests that in a peaceful mechanism, all leaders must ask for the same payoff denoted by $\bar{V}_i := U_i(\bar{\theta}_i)$ no less than $W_i(\bar{\theta})$ *regardless of the audience cost.*

Rewriting citizens' participation constraints is no more difficult: they must always find the peace payoff weakly larger than that of war. This requires that $\forall \theta_i \in [\underline{\theta_i}, \bar{\theta}_i]$,

$$\text{(Citizens)} \qquad W_i^c(\theta_i) \leq X_i(\theta_i)$$

Since a typical citizen of state i benefits from having a strong leader in war, she would expect a war-payoff at most $W_i^c(\bar{\theta}_i)$. Therefore, it costs no more than $W_i^c(\bar{\theta}_i)$ to appease her/him.

Now that we have characterized the properties of an incentive-compatible peaceful menu Σ and domestic stakeholders' participation constraints, it remains to verify whether they jointly ask for more than what the unit-size resource permits. Suppose state i has a leader of type θ_i. At the interim stage, its leader and citizens cannot accept a peace treaty leading to payoffs less than $V_i(\theta_i)$ and $X_i(\theta_i)$. The leader's IC constraint further requires that $V_i(\theta) = \bar{V}_i$. Given the population share assumption, a peaceful settlement would *ex post*

assign state i at least

$$R_i(\theta_i) = \gamma_i \bar{V}_i + (1 - \gamma_i) X_i(\theta_i)$$

for each realization of $\theta_i \in [\underline{\theta_i}, \bar{\theta}_i]$. When a mediator designs the menu Σ, she/he must prepare for the worst situation i.e. two strongest leaders meet at the bargaining table. This is the moment when peace is the most expensive – it costs a total of $\sum_i R_i(\bar{\theta}_i)$. If the unit-size resource is large enough to cover this expense ($\sum_i R(\bar{\theta}_i) \leq 1$), then peace is possible. Substituting in the participation constraints $W_i^c(\theta_i) \leq X_i(\theta_i), W_i(\theta_i) \leq \bar{V}_i$, I derive a necessary condition for the existence of a peaceful menu:

$$\text{(Peace condition)} \qquad \sum_i \gamma_i W_i(\bar{\theta}_i) + (1 - \gamma_i) W_i^c(\bar{\theta}_i) \leq 1 \tag{2.5}$$

I show that Condition (2.5), or the "peace condition", is also sufficient for the existence of peaceful mechanisms.

Result 5. *A peaceful mechanism exists if and only if the peace condition holds.*

Proof. If the condition fails, then for any settlement we can identify a nontrivial subset of Θ that contains at least one unsatisfied stakeholder. To see it, when two strongest types $\bar{\theta}_1, \bar{\theta}_2$ meet, they cannot simultaneously agree on any peaceful settlement simply because doing so would require a budget larger than permitted,

$$\begin{aligned} \sum_i R_i(\bar{\theta}_i) &= \sum_i \gamma_i \bar{V}_i + (1 - \gamma_i) X_i(\bar{\theta}_i) \\ &\geq \sum_i \gamma_i W_i(\bar{\theta}_i) + (1 - \gamma_i) W_i^c(\bar{\theta}_i) > 1 \end{aligned}$$

Since W_i, W_i^c are continuous in θ_i, the set of pairs (θ_1, θ_2) with θ_i close enough to $\bar{\theta}_i$ will for the same reason refuse to sign any peaceful settlement. Because this set of "near-strongest

duos" is non-degenerate, war will happen with strict positive probabilities.

If instead this condition holds, a division (X_1, X_2) satisfying $X_1 \geq W_1(\bar{\theta}_1), X_2 \geq W_2(\bar{\theta}_2)$ would pacify leaders and citizens alike with audience cost setting to zero. □

The characterization of the peace condition should come at no surprise. It generalizes Fey and Ramsay [2011] by taking account of domestic constraints: peace is possible when the total size of resource being divided is large enough to satisfy leaders and citizens in both states. In line with Fey and Ramsay [2011], the peace condition is more benign when the underlying uncertainty is one-sided. It is easily verified that peaceful mechanisms always exist under one-sided uncertainty (cost) due to the war inefficiency[3]. But if uncertainty is modeled as two-sided (relative strength), then peaceful mechanisms exist only when the strongest types of both sides are jointly pessimistic about the expected war outcomes. I defer the formalization to Appendix 2.5.2.

That said, the peace condition is substantively important because it points out the asymmetric roles of audience costs and political bias in conflict resolution. Political bias matters because it affects the minimum payoff necessary to prevent two states from taking the war option. Audience costs do not directly impact the peace possibility, because they essentially describe how a state may redistribute the peaceful settlement between its leader and citizens.

With this observation, we immediately recognize that the (contrapositive of) peace condition provides a sufficient condition under which citizens *would not* use an audience cost that induces peace. This result has implications for the design of optimal domestic constraints in crisis bargaining situations. Specifically, the war technology and political bias have a first-order impact on determining whether peace is materially feasible. If the peace condition fails, then either the leader or citizens of at least one state must wage war with strict positive probabilities regardless of how they may possibly redistribute any peaceful settlement. As such, characterizing the peace condition resolves a theoretical puzzle in Ashworth

3. Take $w_i = p_i - c_i, w_i^c = p_i - \lambda_i c_i$, and note that $p_i + p_j = 1$.

and Ramsay [2017] as to why the war possibility is independent of the domestic preference divergence (shaped by audience costs)[4].

From the peace condition, we can also sharpen a set of robust comparative static predictions known as the "monotonicity results" in the classic game-free analysis [Banks, 1990, Fey and Ramsay, 2011]. In Lemma 7, I deduce from the incentive compatibility condition that stronger leaders must fight (weakly) more often to exploit their military advantage and thereby obtain higher payoffs than weaker leaders. Under the peace condition, a mediator may offer an attractive peaceful settlement to neutralize stronger leaders' excessive war propensity. But in the absence of a peaceful settlement, stronger leaders' excessive war propensity manifests itself again, and the monotonicity result must be strict over some range of types. Formally, taking together the contrapositive of Result 5 and the envelope formula of Lemma 7, we have the strict monotonicity results:

Corollary 1. *If the peace condition fails, then in any equilibrium of the crisis bargaining game there exist strong-type leaders fighting strictly more often than others and obtaining strictly higher payoffs.*

The corollary is a general statement about the properties of equilibria in any crisis bargaining game; therefore, it survives the Banks critique[5]. It identifies sufficient conditions that generate the strict monotonicity results based on model primitives. It advances Banks [1990] who derives a similar result by positing a nontrivial set of leaders fighting with strictly positive probabilities in equilibrium.

4. "Interestingly, the decision about whether or not to use a strategy with positive probability of war is independent of the preference divergence between the leader and the citizen." [Ashworth and Ramsay, 2017, 14]

5. All model predictions "can also be seen as a drawback in that it may be unclear whether the conclusions deduced from a particular model are robust to other specifications of the game." Banks [1990, 599]

Implementation

In this section, I discuss the implementation possibility to complete the game-free analysis. The only interesting case occurs when the resource budget constraint exceeds the minimum required payoff to guarantee domestic stakeholders' participation; otherwise, the peace possibility does not even exist. Hence, we can with loss of generality abstract away domestic politics, take for granted the peace condition, and check whether a mediator may design a procedure to induce peace.

When a mediator has the enforcement power, the extensive-form game Γ is a suitable template for the implementation purpose. According to its procedure, a strong mediator commits to making a proposal (π, x) with $\pi = 0$, followed by two leaders deciding whether to accept the peaceful settlement x. It is easily seen that whenever the peace condition holds, the mediator can propose a division of resource satisfying leaders' participation constraints. Both leaders would indeed accept and honor this peaceful agreement.

When a mediator lacks the enforcement power, implementation becomes very difficult. A key issue is that absent external enforcement, some strong leaders would like to renegotiate the mediated settlement for a higher redistribution of resource at the risk of conflict. Since the standard risk-return trade-off applies again, we may not expect peace to endure. The only way to prevent this from happening is that the leader from other state commits to rejecting any renegotiation possibility, thereby creating the same enforceability environment as before. Whether leaders can make this credible commitment depends crucially on the structure of information uncertainty.

To examine leaders' credibility of commitment, I consider the standard one-sided (fighting cost) and two-sided uncertainty (relative strength) environments à la Fey and Ramsay [2011]. It is easily seen that under one-sided uncertainty environment, the leader with a sufficiently high (private) fighting cost cannot credibly reject all renegotiated offers. But if the uncertainty environment is two-sided, we may construct reasonable off-path beliefs to

back up leaders' credibility of commitment. One viable belief is that only the lowest types would renegotiate and it is rational to trigger war against them. When this is the case, a weak mediator may again implement peaceful settlements according to the procedures in Γ. To summarize,

Result 6. *Strong mediators can implement all possible peaceful settlements. Weak mediators can implement peaceful settlements only when the underlying uncertainty is two-sided, backed by leaders' credible promises to reject any revisionist offer.*

In the Appendix 2.5.2, I formalize when leaders can credibly refuse renegotiated offers.

2.4 Conclusion

This paper analyzes the limits of mediation in general crisis bargaining situations. It mainly addresses three questions: taking account of domestic constraints, when does a peaceful settlement exist? How do different domestic constraints affect the availability of peaceful settlements? Are these settlements implementable by a weak mediator who lacks the enforcement power?

I answer the questions by characterizing the peace condition subject to domestic constraints. With the "game-free" approach, I identify the weighted resource budget constraint as the condition for the existence of peaceful settlements. The characterization indicates that domestic constraints do not fundamentally impact the peace possibility other than affecting the calculus of the resource budget. When the budget constraint fails, stronger-type leaders are more likely to fight and obtain strictly more spoils in crises. I then study how a mediator's enforcement power affects the implementation of peaceful settlements. I show that, while implementation is generally hard without the enforcement power, the mediator may achieve some success when states are uncertain about the relative strength of their adversaries.

My characterization of peaceful mechanisms is a preliminary step to revisit the information rationale for mediation. For mediation to completely preclude any war possibility, it must be that mediators impose settlements acceptable to both sides *without* releasing any payoff-relevant information. Even a weak mediator is potentially helpful, for she could coordinate actions and circumvent wars out of states' greedy ultimatum at the bargaining table. By contrast, if states learn from mediators' offer about the underlying information uncertainty, then their incentive to misrepresent would persist, and war would occur in equilibrium with positive probabilities. Admittedly, failing to completely circumvent war possibility does not mean that mediation is less effective. It just indicates that the information role of mediators depends on the specific mediation institution.

Finally, my characterization serves as a benchmark result for future research on crisis bargaining models. From the contrapositive of the peace condition, I derive a set of strict monotonicity results regarding war propensity and payoff. Since these results hold regardless of the game form and domestic constraints, the value of particular bargaining protocols and domestic constraints lies in generating other empirically-relevant theoretical predictions. I embrace any bargaining details, as long as they are useful towards understanding real-world conflicts.

2.5 Appendix

2.5.1 Proof of Lemma 1

To analyze incentive compatible mechanism, I apply the envelope theorem in Milgrom and Segal [2002]. In an IC mechanism, the leader must be willing to tell the truth. Take the most general case $w_i = w_i(\theta_i, \theta_j), v_i = v_i(\theta_i, \theta_j)$. Since truth-telling is optimal,

$$U_i(\theta_i) = \max_{\hat{\theta}_i} U_i(\hat{\theta}_i|\theta_i) = \max_{\hat{\theta}_i} \int_{\underline{\theta}_j}^{\bar{\theta}_j} [\pi(\hat{\theta}_i, \theta_j) w_i(\theta_i, \theta_j) + (1 - \pi(\hat{\theta}_i, \theta_j)) v(\hat{\theta}_i, \theta_j)] dF_j(\theta_j) \tag{2.6}$$

Since $w_i(\theta_i, \theta_j)$ is differentiable, apply envelope theorem to Equation (2.6)

$$\frac{dU_i(\theta_i)}{d\theta_i} = \int_{\underline{\theta}_j}^{\bar{\theta}_j} \pi(\theta_i, \theta_j) \frac{\partial w_i(\theta_i, \theta_j)}{\partial \theta_i} dF_j(\theta_j)$$

By the fundamental theorem of calculus,

$$U_i(\theta_i) - U_i(\underline{\theta_i}) = \int_{\underline{\theta_i}}^{\theta_i} \int_{\underline{\theta}_j}^{\bar{\theta}_j} \pi(t, \theta_j) \frac{\partial w_i(t, \theta_j)}{\partial t} dF_j(\theta_j) dt \tag{2.7}$$

Equation (2.7) is a necessary condition for incentive compatibility. From standard Bayesian mechanism design argument (see Borgers et al. [2015] for example), we know that it is also sufficient for incentive compatibility if $\bar{\pi}(t) := \int_{\underline{\theta}_j}^{\bar{\theta}_j} \pi(t, \theta_j) \frac{\partial w_i(t, \theta_j)}{\partial t} dF_j(\theta_j)$ is nondecreasing in t.

2.5.2 Credible threat under two-sided uncertainty

Consider the two-sided uncertainty model à la Fey and Ramsay [2011]. Define $\bar{p}_i(t) = \int_{\underline{\theta}_j}^{\bar{\theta}_j} p_i(t, \theta_j) dF_j(\theta_j)$ as state $i's$ expected victory probability with type$-t$ leader and $\bar{c}_i =$

$(1-\gamma_i)c_i + \gamma_i\lambda_i c_i$ the state's average fighting cost. The peace condition (2.5) simplifies to

$$\sum_{k=i,j} \bar{p}_k(\bar{\theta}_k) \leq 1 + \sum_{k=i,j} \bar{c}_k \tag{2.8}$$

as if two unitary actors bargain in the Fey and Ramsay [2011] environment with fighting cost modified to $\bar{c}_k$. As such, peaceful mechanisms exist if and only if when both states are on average pessimistic about war taking account of fighting cost.

When the peace condition (2.5) holds, no one would want to make a revisionist offer if it leads to war with probability one. As such, a large set of peaceful divisions are implementable by a weak mediator, provided that both sides admit a sufficiently weak type who is fought against upon proposing a revisionist offer. The following lemma provides a sufficient criterion to identify the "weak type":

Lemma 8. *If there exists ϵ_i, ϵ_j satisfying the following conditions:*

1. *$p_i(\underline{\theta}_i, \theta_j) \leq \epsilon_i$ for almost all $\theta_j \in \Theta$, $p_j(\theta_i, \underline{\theta}_j) \leq \epsilon_j$ for almost all $\theta_i \in \Theta$.*

2. *$\bar{p}_i(\bar{\theta}_i) \geq \sum_k \bar{c}_k + \epsilon_j, \bar{p}_j(\bar{\theta}_i) \geq \sum_k \bar{c}_k + \epsilon_i$, where $k = \{i, j\}$*

then both leaders' threats to reject any revisionist offer are credible.

Proof. state i can expect at most $1 + \bar{c}_j - \bar{p}_j(\bar{\theta}_j)$ from a peaceful settlement. Facing a revisionist offer, state i expects at least $1-\epsilon_j-\bar{c}_i$ conditional on the belief that the opponent is the worst type. As such, with probability 1 state i would reject a revisionist offer if condition 2 holds. □

How restrictive are the assumptions in Lemma 8? Condition 1 establishes a uniform upper bound for the weakest type's winning probability, which would be trivially satisfied with the Tullock contest success functions (among others[6]) and $\Theta_k = [0, 1]$. Together with

6. See Skaperdas [1996] for a survey.

condition (2.8), there is a large set of parameters consistent with both requirements. So the implementability under two-sided uncertainty is relatively benign.

To study implementation possibility, I posit the following game form: first, a mediator proposes (π, x) with $\pi = 0$. Second, exactly one of the two leaders has a chance to make an ultimatum revisionist offer; each leader has the proposing power with strictly positive probability[7]. War ensues if no agreement is reached.

Under the conditions of Lemma 8, no leader would make a revisionist offer for fear of being perceived as the weakest type. Because renegotiation invites only an inefficient war, on the equilibrium path we observe that the mediator proposes a peaceful settlement that is accepted by both leaders.

7. This assumption is even less restrictive than the "random dictatorship axiom" in Myerson [1984]; it only requires that some leader would be recognized as the proposer in the bargaining game (not necessarily with equal chance).

CHAPTER 3

REASONABLE DOUBT

3.1 Introduction

"Better that ten guilty persons escape than that one innocent suffer".

- William Blackstone, *Commentaries 358*

We encounter testing every day, from electoral competition, law enforcement, to terrorism prevention and judicial decision. Regardless of the context, a near-consensus is that a principal (voter, jury, security agency, etc.) should employ a standard of evidence known as the "reasonable doubt" beyond which to trigger conviction decisions on an agent. To the extent that any test inevitably entails errors, the principal should trade off the social costs of wrongfully convicting the innocent and acquitting the guilty (e.g. Feddersen and Pesendorfer [1998], Posner [1999]).

But testing often concerns more than the principal's strategy, as the classic jury models assume. Anticipating the endogenous standard of the "reasonable doubt", an agent has every incentive to contaminate evidence through costly and hidden efforts. Moreover, the agent must exert efforts according to his/her hidden type. The signaling component differentiates the testing situations from many canonical accountability models in which players are symmetrically informed (see Fearon [1999], Ashworth [2012]).

Two examples illustrate the nature and prevalence of testing. In nondemocracies, office-motivated politicians are less constrained to manipulate GDP statistics. It is hard to think of all politicians choosing the same manipulation activities without knowing their competence. Manipulation activities bear electoral consequences: high-offices judge subordinates by those tangible criteria. Utilizing satellite data to match nightlights with government statistics, Martinez [2019] finds that officials in nondemocratic regimes are particularly prone to exaggerate their GDP statistics before elections. But GDP manipulation does not necessarily

come for free. It may not only divert politicians' attention from delivering constituency service, but also cost their reputation and deteriorate their long-term relationship from superiors in charge of this jurisdiction [Jiang and Wallace, 2017].

In democratic regimes today, politicians have to expend more efforts safeguarding their valuable but vulnerable reputation against the challenges of new information technology. In 2019, an article titled "*Deepfakes and the New Disinformation War*" in the *Foreign Affairs* magazine warns us of the danger that the "deepfake" technology imposes on democratic norms. Deepfakes refer to synthetic media that are highly realistic and difficult to detect. Once well-timed, they may have the potential to tip elections. According to this article, Emmanuel Macron nearly fell victim of the Russian hackers who attempted to undermine his 2017 presidential campaign with forgery documents. To guard against the deepfakes, the authors recommend precautionary measures such as alibi service to the high-profiles, but concede that we may learn to "live with lies". It remains ambiguous whether the constituents would *perceive* the politicians as innocent and cast their votes, anticipating necessary measures taken by politicians against the disinformation war.

In this paper, I study how a principal interacts with an agent in a testing situation. I ask: what does the equilibrium strategy profile look like? If a principal sets a reasonable doubt, how should she maintain her doubt reasonable subject to changes in the testing environment?

To answer these questions, I develop a model of testing based on the standard signaling model of political agency. The principal-agent relationship of the testing game naturally fits the jury context: a judge or a pivotal juror wants to learn the type of a defendant from the result of an innocence test. It also captures the strategic interactions of political selection as in my motivating examples: the constituency wants to select the high-competence politician based on whether he/she "appears good". In testing, the principal has to choose from the set of binary actions labeled "convict" or "acquit" on the basis of a test standard unknown to the agent. For example, there may lack definitive promotion criteria for officials in nondemocratic

regimes; the constituency may lack the technical expertise to consistently tell true scandals from the deepfakes.

Key to the testing situation is the agent's influence on the test result based on his conjecture of the principal's standard. He does so by exerting costly and unobserved efforts according to his type. For example, an outside observer never knows how carefully politicians manage to insulate themselves from scandals. I assume that the test result is determined by the agent's type and effort, plus an idiosyncratic shock that has a Gaussian distribution. This assumption is standard in the career concern models à la Holmström [1999]. That said, our strategic nature are quite different because of the signaling component. In my model, an innocent agent is always the better test-taker than a guilty one simply due to his innocence. His superior testing technology may take the form of an *effort advantage*, if he expects better test results by expending the same (positive) effort. The innocent agent may also possess an *initial advantage*, if he tends to outperform when both types exert zero effort. I take the effort advantage for granted, but dichotomize tests with respect to whether the initial advantage exists.

Specifically, I call a testing situation *criminal* whenever the innocent agent possesses the initial advantage, and *civil* otherwise. Motivated by the real-world judicial decision-making, this terminology reflects the following idea: unfavorable evidence about a (truly) guilty agent is more likely to be discovered in criminal investigations than civil ones. To see it, fix all agents' efforts to the same level. A judge in a civil test is equally likely to discover unfavorable evidence (signals, test results, etc) for both types of agents; in a criminal test, she is much more likely to discover unfavorable evidence about a guilty agent relative to an innocent one. Put differently, a guilty agent finds it much harder to disguise himself as innocent in criminal tests.

Main results. I establish the analytical equivalence between the testing game and a Bayesian inference problem. The argument follows from two crucial observations. First, the

principal must employ a threshold conviction strategy in spirit of the "reasonable doubt" in *any* testing equilibrium. The reason is that better results are more likely to come from the innocent type since he is the better test-taker. Hence, the principal should order the plausibility of innocence solely based on test results, and find a correct threshold (test difficulty) to optimally trade off two-type inference errors. Second, I show that the principal can do so by computing the value of a statistic known as the "test informativeness". This statistic derives from the likelihood ratio of the test, but it accounts for the agent's best response for each test difficulty. This means that the principal can carefully control the test informativeness as if it were her own decision problem. If the informativeness of a particular test matches the relative harm of two-type errors – also known as the *Blackstone ratio*[1] – then its associated difficulty level is a candidate solution for the principal. I prove that any match indeed corresponds to a testing equilibrium: fixing the agent's anticipated response, any other test difficulty would deem too hard or too easy from the principal's ideal level.

I also examine how the equilibrium test difficulty responds to changes in the testing environment. The key lemma is that, a criminal (civil) test tends to be more (less) informative as its difficulty goes to the extreme. Here is the intuition: in an extremely hard/easy test, the agent expects an almost-sure conviction/acquittal decision and therefore is not very motivated to exert effort. In this case, an innocent agent stands out from a guilty one if and only if he possesses the initial advantage. This observation alludes that we may have opposing comparative static predictions in civil and criminal tests respectively. For example, suppose the principal demands an uninformative test; it happens when she wants to emphasize guilty conviction, but still believes that innocent protection is her job priority. Then we may observe an easier civil test or a harder criminal test in equilibrium.

I conclude the introduction by relating my work to the existing literature. Focusing on

1. "*All presumptive result of felony should be admitted cautiously; for the law holds it better that ten guilty persons escape, than that one innocent party suffer*" [Blackstone, 1962]. It is a succinct way of measuring the social attitude towards acquitting a guilty vis-a-vis convicting an innocent.

the moral hazard aspect of testing, my paper complements the formal studies of the Condorcet jury theorem and committee design [Condorcet, 1785, Feddersen and Pesendorfer, 1998, Coughlan, 2000, Persico, 2004, Gerardi and Yariv, 2008]. A near-axiom is that information aggregation in collective decision making could fail because each juror responds strategically to their (binary) signals. As such, there is room for improving the design of voting mechanisms, either by allowing pre-voting communication or motivating information acquisition. Instead of studying the strategic interactions among jurors, my model looks at the defendant's "supply side" of information and works with a rich signal space. For a jury interpretation, my model disentangles how a pivotal juror who votes sincerely might choose a threshold of reasonable doubt that maximizes the *accuracy* of her signal.

My paper relates to the optimal design and enforcement of law. Common in the theme is the social planner's trade-off between crime deterrence and the enforcement cost (as in Becker [1968] and Stigler [1970]). To delve deeper into the supply side of crimes, later works tackles the deterrence problem with a mechanism design approach where the suspects choose their activities in an incentive compatible way [Mookherjee and Png, 1994, Kaplow, 2011, 2017]. Among them, Kaplow [2011] is the closest to mine. He extends the deterrence framework by endogenizing the conviction threshold that inevitably induces the errors of convicting the innocent and acquitting the guilty. Our approaches differ in two crucial aspects. First, I assume that testing is not costly. Instead of doing the cost-benefit analysis of deterrence, my model concerns how the principal might possibly make informed decisions. Second, I relax the principal's power to commit to her test standard as required by the mechanism design approach. The no commitment assumption is natural in the jury context, but it is also plausible in many judicial decision-making situations. For instance, legal doctrines often entail purposeful vagueness due to court's hierarchic structure [Lax, 2012]. Consequently, agents do not necessarily observe the exact test standard before putting their hidden efforts.

My paper also contributes to the political agency literature. The testing model concerns

the "pure selection" aspect of political agency, which differentiates itself from the "pure hazard" retrospective voting models [Barro, 1973, Ferejohn, 1986] and the "career concern" models [Holmström, 1999, Fearon, 1999, Besley, 2006, Ashworth, 2012, Ashworth et al., 2017a,b]. In those models, a political agent unaware of her competence exerts effort to improve the reelection prospect. If instead the agent reacts to competence, her rich signaling strategy would complicate the analysis of the model. Viewing this, my paper contributes to the literature by providing a computational toolkit based on the analytical equivalence.

The rest of the paper is organized as follows: in Section 3.2, I present the structure of the testing game. I analyze the game in Section 3.3, and discuss its implication in Section 3.4. Section 3.5 concludes.

3.2 Model

Setup. Two players, a principal ("she") and an agent ("he"), play a testing game. The agent has a hidden "type" $\theta \in \Theta := \{I, G\}$ indicating whether he is **I**nnocent or **G**uilty. Order Θ as $I \succ G$, meaning that the innocent is the "higher type". The principal wants to learn the type of the agent with a test. The test produces an observable result $s \in S$, upon which the principal chooses one of the binary decisions. Formally, the principal's strategy is $\sigma_p : S \to D := \{\mathcal{C}, \mathcal{A}\}$ mapping from the test result to actions "$\mathcal{C}$onvict" or "$\mathcal{A}$cquit". Since the state of the world Θ is binary, we can assume $s \in S = \mathbb{R} \cup \{\pm\infty\}$, with higher s being more indicative of guilty. Let $\mathcal{G}$ denote the (measurable) set of all test results at which the conviction decision would be triggered. As such, we may also think of the principal's strategy as choosing the set $\mathcal{G} \subset S$.

The agent in the test chooses an effort e corresponding to her type θ. Formally, his strategy is $e : \Theta \to \mathbb{R}_+$. His test result is generated according to a function of type, effort,

and white noise:

$$s = -h(\theta, e) + \epsilon$$

$h(\theta, e)$ can be thought of the type-θ agent's "expected performance" in the test. It is non-negative, strictly increasing and differentiable in his effort e. This specification implies that more efforts tend to reduce s, which makes the agent appear more innocent. Denote $h_0 := h(I, 0) - h(G, 0)$. Because the innocent agent is the better test-taker, $h_0 \geq 0$. If $h_0 > 0$, we say that the innocent agent possesses an *initial advantage*. Since h is monotonically increasing in e, we may think of the agent's strategy as directly choosing his "expected performance" h instead of effort e. This means that I can write type-θ agent's strategy as $h_\theta := h(\theta, e(\theta))$ whenever it is not necessary to emphasize the role of efforts explicitly.

The cost associated with the expected performance h is $C_\theta(h)$ for a type-θ agent. For $\theta \in \{I, G\}$, $C_\theta(h)$ is smooth and increasing, and $C_\theta(h(\theta, 0)) = C'_\theta(h(\theta, 0)) = 0; C''_\theta(h) > 0, C'''_\theta(h) \geq 0$ for all $h > h(\theta, 0)$. Here $h(\theta, 0)$ is the test result that a type-θ agent can get for free. The condition restricts the shape of the cost function. It basically says that better test results are increasingly costly.

Finally, the white noise ϵ also impacts the test result. ϵ is independent of action or type; it can stand for the malfunction of the test machine, or the principal's imprecise interpretation of the test results. I assume that ϵ is distributed according to standard Gaussian distribution. Its probability density function (pdf) ϕ is atomless, single-peaked, symmetric around 0 with unbounded support. Furthermore, ϕ has the *monotone likelihood ratio property (MLRP)*: the likelihood ratio $\mathcal{L}(s, h_I, h_c) = \frac{\phi(s+h_I)}{\phi(s+h_G)}$ is strictly *decreasing* in s for all $h_I > h_G$.

To recap, the sequence of the game is as follows: (1) Nature draws the agent's type $\theta \in \{I, G\}$. (2) The agent chooses his "expected performance" h based on his type. (3) The test result realizes and the principal decides whether to convict the agent.

Preferences. All players in the game are risk neutral.

The agent cares about both the sufferings from conviction decisions and the cost associated with improving test performances. I normalize the punishment from the conviction decision to 1. Because the binary decisions $\{\mathcal{A}, \mathcal{C}\}$ are functions of players' strategy profile, I write the agent's objective as

$$\max_{h_\theta} u_\theta(\sigma_p, h_\theta, h_{-\theta}) = -\mathbb{P}\{\mathcal{C}(\sigma_p, h_\theta, h_{-\theta})\} - C_\theta(h_\theta)$$

The principal cares about minimizing the social cost of making wrong decisions. Her prior belief that an agent is innocent is λ. Wrongfully acquitting a guilty costs her $q \in (0,1)$, while wrongfully convicting an innocent costs her $1-q$. Her expected disutility from making wrong decisions is

$$q(1-\lambda)\mathbb{P}\{\mathcal{A}(\sigma_p, h_I, h_G)|\theta = G\} + (1-q)\lambda\mathbb{P}\{\mathcal{C}(\sigma_p, h_I, h_G)|\theta = I\}$$

By defining $\alpha = \frac{q(1-\lambda)}{\lambda(1-q)}$, I reformulate the principal's objective as

$$\begin{aligned}
\max_{\sigma_p} u_p(\sigma_p, h_I, h_G) &= -[\mathbb{P}\{\mathcal{C}(\sigma_p, h_I, h_G)|\theta = I\} + \alpha\mathbb{P}\{\mathcal{A}(\sigma_p, h_I, h_G)|\theta = G\}] \\
&= [1 - \mathbb{P}\{\mathcal{C}(\sigma_p, h_I, h_G)|\theta = I\}] - \alpha\mathbb{P}\{\mathcal{A}(\sigma_p, h_I, h_G)|\theta = G\} - 1 \\
&= \mathbb{P}\{\mathcal{A}(\sigma_p, h_I, h_G)|\theta = I\} - \alpha\mathbb{P}\{\mathcal{A}(\sigma_p, h_I, h_G)|\theta = G\} - 1
\end{aligned}$$

From this expression, we may interpret the principal's objective as the maximizing the weighted acquittal probability difference of the innocent and the guilty. The weight is given by the *Blackstone ratio* α, which reflects a society's relative tolerance towards two-type inference mistakes.

Technical assumptions. I impose that the Blackstone ratio $\alpha \in (0,1)$. The restriction

is essentially a *consistency* requirement: the principal *always* prioritizes on dealing with the mistake of wrongfully convicting the innocent. In his original quote "*Better that ten guilty persons escape than that one innocent suffer*", Sir Blackstone expresses his ideal rule that $\alpha \leq \frac{1}{10}$. The restriction is harmless with respect to the equilibrium characterization. When it comes to comparative static analysis, the restriction guarantees that we are comparing the (selected) equilibrium outcomes with respect to parameter changes in a meaningful way.

I assume that the innocent agent is the better test-taker measured in terms of cost. Specifically, the cost function $C_\theta(h)$ satisfies the strict Spence-Mirrlees Property (SMP): $C'_G(h) > C'_I(h)$ for all $h \in [h(\theta, 0), \infty)$.

I also impose the following selection rule to ensure that the agent has a well-defined best reply,:

Assumption 3 (Selection rule)**.** *At least one of the following is true:*

1. $C''_\theta(h) > \int_{\mathbb{R}} |\phi''(x)|dx$ *for all* $h \geq h(\theta, 0)$.
2. *The type-*θ *agent selects on the largest or smallest* h^*_θ *from the best response correspondence* $\{h^*_\theta(\sigma_p)\}$.

Condition 1 simply restricts the curvature of cost function, which guarantees the concavity of the agent's maximization program. It is well-defined because $\int_{\mathbb{R}} |\phi''(x)|dx \leq 1 + \frac{1}{2\sqrt{2}}$. Since Condition 1 guarantees the uniqueness of the agent's best response, it implies Condition 2 trivially. Condition 2 says that conditional on the principal's strategy σ_p, if both types of the agent have more than one best reply, then they will use the same rule in selecting elements from the set of maximizer $\{h^*_\theta(\sigma_p)\}$. The condition is common in the literature of monotone methods; it basically makes the agent's efforts comparable across types.

Solution concept. The solution concept of the testing game is Perfect Bayesian Equilibrium (PBE). This is because the game moves sequentially and the agent's type is unknown.

Once the agent makes his effort input, the principal forms beliefs about the agent's innocence from the test result. Since s has full support, there is no off-path belief. Furthermore, because the game does not necessarily admit a unique solution, I adopt the equilibrium selection that the principal uses the most lenient (harshest) conviction strategy in a civil (criminal) test. We may microfound this selection criterion as the principal's lexicographic preference for the maximal leniency (deterrence) in civil (criminal) tests. Since the principal's equilibrium belief is implied by the equilibrium strategy profile, I omit the description of belief and hereafter simply call any PBE "equilibrium".

Comments. Before analyzing the model, I make the following comments:

Deterrence. This model does not study the deterrence effects of the principal's conviction strategy as in the classic economic approach of law. Key to the deterrence effects is the agent's fear of punishment (imposed by the principal) that deters him from conducting harmful deeds. For instance, Becker [1968] and Stigler [1970] debate whether the extreme sentence is good for crime deterrence; Lando [2006] analyzes whether wrongful convictions may achieve crime deterrence. While in my model the principal's strategy has an influence on the agent's action, she cannot credibly threaten to impose any arbitrary conviction strategy to deter the guilty agent's harmful deeds. The endogenous nature of her strategy makes its influence different from what is perceived as deterrence.

Decision cost. For the testing problem to be nontrivial, I abstract away the principal's decision cost. The assumption is realistic in the context of political selection. For instance, it never costs the electorates anything if they want to pick a higher retention threshold for the incumbent. Introducing decision cost would blur the inference nature of testing. Suppose instead the principal pays a convex cost for increasingly "harder" tests, then essentially she trades off better tests and higher testing expenses. This theme has been fruitfully addressed by the law and economics literature and Ting [2017].

Initial advantage. Whether the innocent has the initial advantage depends on the testing context. It concerns whether an innocent agent distinguishes from a guilty one by his innocent type or type-induced effort. It seems reasonable to think of the innocent as possessing an initial advantage in criminal cases. For example, during murder investigations, a truly-innocent suspect can easily separate from a guilty one if neither contaminates evidences. Stories are different in civil or preventive activities, such as airport security checks or online content censorship. Travelers frequently encounter additional security checks when they forget to empty the pockets. An online article[2] suggests that poor censorship technology may put innocent netizens on the wrong side when they post anti-terrorism images or slogans.

The disinformation war seems to belong to civil tests. The reason is that, as highlighted by the Macron example, the disinformation war often targets at unprepared politicians with fake news[3]. Since neither truly innocent (clean) or guilty (corrupted) politicians may timely refute the charges, they face the same trouble and suffer alike from the disinformation war. Along this line, the authors of the *Foreign Affairs* magazine attribute Macron's bare escape from the deepfakes not to his innocent nature, but the protection of the French media law.

Bayesian inferences. The principal's optimization problem also has a Bayesian interpretation. Before the test, the principal assigns λ as the prior belief that the agent is innocent. The prior is not good enough for the principal's cost-benefit analysis, so she needs to refine her belief by incorporating new evidence from the test. Her job is to carefully design the difficulty of the test without observing the data, and define "good news" and "bad news" upon which to trigger conviction decisions. The posterior determined by the equilibrium likelihood ratio must achieve the principal's desired tradeoff in balancing two-type errors.

2. "Industry Efforts to Censor Pro-Terrorism Online Content Pose Risks to Free Speech". Last modified July 12, 2017. Electronic Frontier Foundation.

3. Different from the traditional negative campaign, the disinformation war does not have to ground on verifiable record. See Polborn and Yi [2006], Egorov [2015] for formal analyses of negative campaigns.

3.3 Analysis

The strategy profile $(\mathcal{G}^*, h_\theta^*)$ constitutes an equilibrium if and only if 1) players successfully coordinate on their mutual best responses 2) players do not admit any profitable unilateral deviation from the conjectured profile.

The main technical difficulty of equilibrium characterization concerns how to think of the principal's optimal strategy. It is a priori unclear whether a principal may gain by unilaterally choosing a standard $\mathcal{G}'$ different from the conjectured strategy $\mathcal{G}^*$.

To overcome the difficulties, I prove a few regularities that are valid in *any* equilibrium. I show that we may analyze the game as if solving an inference problem from the principal's perspective. To keep track of the main ideas, below I provide the intuition and sketches of the main results, and defer the formal analysis to the appendix.

3.3.1 Principal's cutoff strategy.

The crucial observation is that, in any testing equilibrium, the principal must employ a cutoff strategy in spirit of the "reasonable doubt". The argument proceeds in two steps.

On average, the innocent agent performs better in any test. The result follows immediately from the innocent agent's superior testing technology. Since he achieves any expected test score h at a lower marginal cost, the innocent agent must have a higher expected performance than a guilty type.

Recognizing this, the principal sets a threshold and convicts if the agent's performance falls short of it. This is because, after applying the Bayes's rule, the principal concludes that better test scores are more likely to come from an innocent agent. Put differently, better test scores are indeed better news of being truly innocent. Thus, the principal may choose at most one cutoff s^* to dichotomize the agent's performance.

It is worth noting that in equilibrium, the principal may employ degenerate cutoff strategies i.e. acquitting or convicting at all test results. To accommodate this case, I call a testing

equilibrium (s^*, h_θ^*) *trivial* if $s^* = \pm\infty$ and *nontrivial* otherwise. The rest of my analysis devotes to characterizing the conditions under which nontrivial equilibria exist.

3.3.2 Equivalence to correct inference.

I simplify the players' objectives as follows. Since the principal's optimal strategy must take a cutoff form, she chooses a threshold s^* and convicts whenever $s \geq s^*$ (because higher s is bad news). This means that the agent is acquitted with probability $\mathbb{P}(s < s^*) = \mathbb{P}(\epsilon < s^* + h) = \Phi(s^* + h)$. Clearly, smaller s^* means harder tests. After relabeling, I rewrite players' objectives in terms of the normal CDF Φ,

$$\textbf{Agent} \qquad \max_h u_\theta(s, h) = \Phi(s + h) - C_\theta(h), \quad \theta \in \{I, G\}$$

$$\textbf{Principal} \qquad \max_s u_p(s, h_I, h_G) = \Phi(s + h_I) - \alpha\Phi(s + h_G)$$

The first-order conditions of the optimization program implies a set of well-defined best replies. For the agent, he optimizes over the set of influence activities h based on his conjecture about the principal's threshold $\hat{s}$. Write the agent's best reply to $\hat{s}$ as $h_\theta^*(\hat{s})$. Given $h_\theta^*(\hat{s})$, the principal must choose s^* to solve

$$\left.\frac{\phi(s^* + h_I^*(\hat{s}))}{\phi(s^* + h_G^*(\hat{s}))}\right|_{\hat{s}=s^*} = \alpha$$

This condition is both necessary and sufficient for the principal's optimization program in any (nontrivial) equilibrium profile. Thanks to the MLRP of the white noise, the principal's objective is single-peaked at s^*. The intuition is that, having fixed the agent's conjectured behavior $h_\theta^*(\hat{s})$, the principal trades off two-type inference errors at a rate that is decreasing in s. At the equilibrium level s^*, which is correctly anticipated by the agent (who sets $\hat{s} = s^*$), the principal achieves her desired trade-off – the Blackstone ratio α. She does not want to unilaterally deviate from the conjectured equilibrium (s^*, h_θ^*), for otherwise she

would make the test too harsh or too lenient.

The above discussion identifies the function $LR(s) := \frac{\phi(s+h_I^*(s))}{\phi(s+h_G^*(s))}$ as a key statistic of the testing game. $LR(s)$ resembles the likelihood ratio $\mathcal{L}(s, h_I, h_c)$ in that it captures how *informative* the test is. For example, if $LR(s)$ is close to 1, then the principal cannot accurately tell the innocent from the guilty type at the difficulty level s. That said, no assumption insures the monotonicity of $LR(s)$. This means that the solution to the equation $LR(s) = \alpha$ does not have to be unique. But whenever there exists an s^* such that $LR(s^*) = \alpha$, then we may recover an equilibrium profile (s^*, h_θ^*) of the testing game with $h_\theta^* = h_\theta^*(s^*)$.

I formally state the equivalence result:

Proposition 3 (Equivalence to an inference problem)**.** *The testing game is analytically equivalent to the principal's statistical inference problem:* (s^*, h_θ^*) *is an equilibrium if and only if* $LR(s^*) = \alpha$.

Proposition 3 suggests that the testing game may exhibit equilibrium multiplicity. This conclusion holds because more than one s may solve $LR(s) = \alpha$. Alternatively, we may interpret this multiplicity as a natural consequence of coordination. In testing, neither the principal nor the agent can observe the others' action before choosing their own. This means that players have to coordinate on some mutually-correct conjecture (s^*, h_θ^*). Successful coordination does not have to be unique (as in the canonical battle of the sexes game).

Remark. Matters are entirely different if the principal commits to a testing threshold s that is observable to the agent. Specifically, the principal can no longer play the testing game as if solving a statistical decision problem. The reason is as follows: if the agent observes s before taking his action, then his strategy maps from type *and* test threshold to his influence activities. This means that if the principal deviates from some conjectured level s^*, then the agent would make appropriate strategic adjustments. In other words, the principal's deviation is *anticipated*. The agent's response makes the principal's optimization problem neither concave nor single-peaked ex ante. As such, the set of first-order conditions no longer

suffices to characterize the testing equilibrium.

3.3.3 Analyzing test informativeness.

In light of the equivalence result, I study the test informativeness $LR(s)$ and check if any s induces the desired informativeness α. The following lemma is the key step:

Lemma 9. *The agent's effort incentive is not monotone with respect to the test difficulty.*

Here is the intuition. When the test is moderately difficult, the agent is very motivated because his (marginal) effort has a high return in terms of reducing the conviction probability. For example, if the agent's performance nearly offsets the test difficulty ($s + h \approx 0$), then the agent benefits the most ($\phi(0)$) by exerting additional efforts. By contrast, the agent is discouraged to exert effort in an extremely easy/difficult test. In this case, his additional effort has little chance to influence the principal's almost-sure decision. Therefore, the agent prepares the most for a moderately difficult innocence test.

We are now in a good position to study the properties of the statistic $LR(s)$. By construction, it is continuous and differentiable in s. In what follows, I consider its *interior behavior* ($s \neq \pm\infty$) and *asymptotic behavior* ($\lim_{s\uparrow\infty} LR(s)$) respectively.

Interior behavior. The key interior property of $LR(s)$ is that, there exists an interval $[\underline{s}, \bar{s}]$ in which $LR(s)$ decreases from above to below one. The logic is as follows. For each type-θ agent, there exists a unique test difficulty s_θ at which he is most motivated to work i.e. $s_\theta + h^*_\theta(s_\theta) = 0$. Since the innocent type is the better test-taker, $h^*_I(s) > h^*_G(s)$ for any s. Thus, it must be that $s_I < s_G$. Note also that $LR(s_I) = \frac{\phi(s_I + h^*_I(s_I))}{\phi(s_I + h^*_G(s_I))} = \frac{\phi(0)}{\phi(s_I + h^*_G(s_I))} > 1$ because the normal pdf ϕ is single-peaked at 0. Similarly, $LR(s_G) < 1$. We can prove that $LR(s)$ is decreasing for $s \in [s_I, s_G]$. The reason is that, test difficulties of this level are moderate for the innocent but overwhelmingly high for the guilty type. Consequently, an increase in test difficulty ($s \downarrow$) would elicit more efforts from the innocent but inhibit efforts

from the guilty type. Taken together, this leads to a higher $LR(s)$. To conclude, we have identified the desired interval $[\underline{s}, \bar{s}]$ with $\underline{s} = s_I$ and $\bar{s} = s_G$.

Two consequences follow immediately. First, $LR(s) > 1$ for all $s \leq \underline{s}$. This is because $s + h_G^*(s) < s + h_I^*(s) < 0$ for all $s \leq \underline{s}$ and ϕ is single-peaked at 0. Likewise, $LR(s) < 1$ for all $s \geq \bar{s}$. Second, there exists an $s_1 \in [\underline{s}, \bar{s}]$ with $LR(s_1) = 1$ by the continuity of $LR(s)$. Together with the restriction that $\alpha \in (0, 1)$, we can without loss focus on the interval $[s_1, \infty]$ to analyze the equilibrium existence properties[4].

Asymptotic behavior. I show that $\lim_{s\uparrow\infty} LR(s) = 1$ if $h_0 = 0$ and $\lim_{s\uparrow\infty} LR(s) = 0$ if $h_0 \neq 0$. In words, the innocent stands out from a guilty one in an extremely easy test if and only if he possesses the initial advantage. To see this, an extremely easy test would heavily discount the innocent agent's effort advantage. Absent the initial advantage, both types of agent would perform almost equally well, thereby making $LR(s)$ close to 1. But if he indeed possesses the initial advantage, an innocent type would immediately distinguish himself and make $LR(s)$ close to 0. Since the initial advantage exists only in criminal tests,

Lemma 10. *The civil (criminal) test becomes more uninformative (informative) as its difficulty goes to the extreme.*

Now let us collect the properties of $LR(s)$ in different tests. In civil tests, $LR(s)$ equals 1 at s_1 and $+\infty$ in the extended real line. Its continuity property suggests that $LR(s)$ would achieve its minimum at some $s_0 \in (s_1, \infty)$. Therefore, in civil tests $LR(s)$ has a range $[LR(s_0), 1]$ because it is bounded above by 1. In criminal tests, $LR(s)$ equals 1 at s_1 and 0 at $+\infty$. This means that $LR(s)$ is onto $(0, 1)$ for $s \in (s_1, \infty)$.

I illustrate these facts by plotting the $LR(s)$ of a typical test with $C_I(h) = \frac{h^2}{4}$ and $C_G(h) = \frac{h^2}{2}$ in Figure 1. This is a civil test: without costly efforts ($C_I = C_G = 0$) no agent would stand out in the test. Extremely easy ($s \approx 3$) or extremely difficult ($s \approx -3$) tests are

4. This does not mean that we should make this restriction on α in any test. But with the technique developed here, we may easily extend the analysis to other situations.

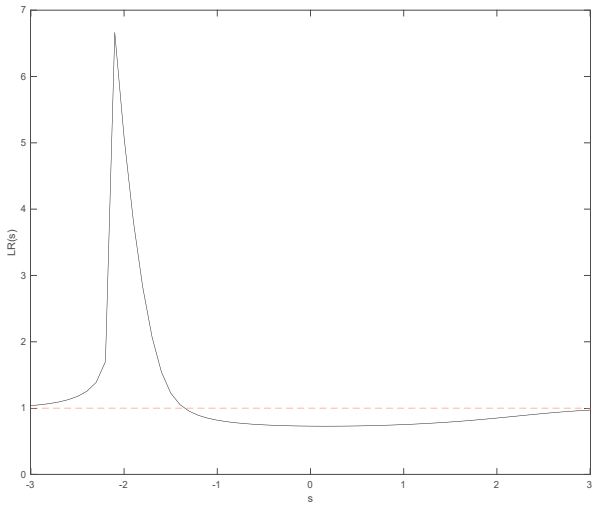

Figure 3.1: The function $LR(s)$ with $C_I(h) = \frac{h^2}{4}$ and $C_G(h) = \frac{h^2}{2}$.

uninformative ($LR \approx 1$); moderately difficult tests (say, $s \approx -2$) are much more informative.

3.3.4 Closing the model.

At this point, we are ready to state the main result.

Proposition 4 (Existence)**.** *A testing equilibrium exists. In particular,*

1. *in civil tests* ($h_0 = 0$)*, a nontrivial testing equilibrium exists if and only if* $LR(s_0) \leq \alpha < 1$*;*

2. *in criminal tests* ($h_0 > 0$)*, a nontrivial testing equilibrium always exists.*

The logic of this proposition is as follows. If α falls into the range of $LR(s)$, then by the intermediate value theorem (IVT) we conclude the existence of an s^* (not necessarily unique) such that $LR(s^*) = \alpha$. Otherwise, the principal has to let go all suspects for fear of wrongfully convicting too many innocent agents.

It is worth mentioning that in civil tests, the nontrivial equilibrium is generically not unique. To see this, consider how the principal may design a test more uninformative than s_0.

She may make it harder (s closer to s_1) or easier (s closer to infinity) than s_0; either way, the new test would discount the innocent type's effort advantage and force closer performances across types. The non-uniqueness property may not hold in criminal tests. The reason is that, while as before the principal can reduce the test informativeness by making it harder, she may not do the same with an easier test. In criminal tests, an extremely easy test tends to be very informative.

The above discussion suggests that, after positing different equilibrium selection rules in different tests, we may observe opposing equilibrium test difficulty changes with respect to changes in α.

3.4 Implications

To illustrate the real-world relevance of the testing model, now I perform some comparative static analysis. Before proceeding, however, it is important to come back to the notion of the Blackstone ratio α. Here are two interpretations.

Following its original formulation, we interpret α as how much a principal is willing to trade off two-type inference errors. If α increases, it means that she no longer considers the mistake of wrongfully convicting an innocent as severe as before.

In the context of electoral tests, we may alternatively interpret α as a politician's incumbency (dis)advantage[5]. Intuitively, the reciprocal of α measures how expensive it is for the electorate to replace the incumbent. At one extreme ($\alpha = 1$), the electorate considers the incumbent *a priori* identical as a potential challenger in terms of competence or innocence. At the other ($\alpha \approx 0$), the electorate wants to retain the incumbent unless there is overwhelming negative evidence.

5. To see it, let λ, λ_c be the prior belief that an incumbent/challenger is clean. The Bayes rule suggests that the constituency would like to retain the incumbent if her posterior belief after the test satisfies $\frac{\lambda\phi_I}{\lambda\phi_I+(1-\lambda)\phi_G} \geq \lambda_c$, or $LR := \frac{\phi_I}{\phi_G} \geq \frac{\lambda_c}{1-\lambda_c}\frac{1-\lambda}{\lambda}$. Define the Blackstone ratio $\alpha := \frac{\lambda_c}{1-\lambda_c}\frac{1-\lambda}{\lambda}$. The incumbent may plausibly have an advantage ($\alpha \in (0,1)$): the fact that the incumbent stays in office indicates that he/she is perceived more "decent" than a potential challenger ($\lambda > \lambda_c$).

Either way, increasing α from 0 to 1 suggests that the principal desires a more uninformative test. Per discussions about the properties of $LR(s)$, the test nature matters when we interrogate the empirical implications of the model.

3.4.1 Blackstone ratio.

Suppose a judge/jury seeks to place more emphasis on convicting the guilty, though acquitting the innocent is still her priority. How should we predict changes in the equilibrium conviction probability?

At first glance, the answer seems unambiguous. It is reasonable to expect more equilibrium convictions, since wrongfully convicting the innocent becomes the lesser of two evils for the principal. This intuition indeed carries through, for example, in the context of strategic voting [Feddersen and Pesendorfer, 1998].

But the intuition is only partially true in testing situations. Perhaps surprisingly, we shall observe higher equilibrium conviction probabilities in *criminal* tests, but lower in *civil* tests.

The result hinges crucially on how informatively the principal wants to tell an innocent agent from a guilty one. Fix any $\alpha \in (0, 1)$. A higher α means that the principal desires a more *uninformative* test in equilibrium. In civil tests, she can make the test easier to mute an innocent agent's advantage over a guilty one; in criminal tests, she can do the opposite thing: by making the test harder, the principal moves a step closer to the most uninformative test (s_1), and thereby achieves the maximal possible deterrence.

The link between test difficulty and conviction probability is straightforward. In the appendix, I show that the test difficulty always overwhelms the agents' effort. Therefore, harder tests always translate to higher conviction probabilities.

Put together all the discussion,

Result 7. *The equilibrium test difficulty and conviction probability is*

- *increasing in α in criminal tests;*
- *decreasing in α in civil tests.*

Remark. It is worth mentioning the role of the restriction $\alpha \in (0,1)$ in this comparative static exercise. As a measure of test informativeness, restricting $\alpha \in (0,1)$ allows us to rank tests in a monotonic way. To see this, consider the following extreme example: increasing α from 0 to 1 suggests that the principal does not want to distinguish two types of agents. But increasing α from 1 to ∞ suggests that she cares about distinguishing agents again. Positing $\alpha \in (0,1)$ also reflects substantive concerns in a number of testing situations. For example, Sir Blackstone's ideal judicial practice restricts $\alpha < \frac{1}{10}$; wrongfully convicting the innocent outweighs acquitting the innocent in terms of the social cost; and an incumbent politician often possesses some incumbency advantage.

3.4.2 Testing Technology.

Suppose the guilty agent has improved his testing technology. Specifically, let's assume that for each (mean) test score h, the guilty agent pays less than he used to, but still more than the innocent. In equilibrium, shall we expect a harder or easier test?

We may reasonably expect a harder test and a higher conviction probability in equilibrium. Intuitively, better testing technology enables the guilty to perform better and appears more similar to the innocent. This urges the principal to tighten up the test standard in order to differentiate agents.

But this intuition is incomplete. Let's again contemplate what happens to the test informativeness subject to the technological change. Fix the test difficulty at the previous level s^*. The crucial observation is that, even if the guilty agent's performance improves from $h_G^*(s^*)$ to a higher $\tilde{h}_G^*(s^*)$, the new test informativeness measured by the likelihood ratio may increase or decrease. To see this, recall that the normal pdf $\phi(x)$ is single-peaked at 0; it is decreasing whenever $x \geq 0$. Together, this means that unless $s^* + h_G^*(s^*) \geq 0$, in

which case the likelihood ratio surely goes up if $\tilde{h}^*_G(s^*)$ replaces $h^*_G(s^*)$, (in all other cases) we cannot determine the direction of changes in the test informativeness. Accordingly, we may not predict equilibrium test threshold changes without calibrating model parameters.

Luckily, we can say sure thing about civil tests. I prove in the appendix that in civil tests, a nontrivial equilibrium admits the property that $s^* + h^*_\theta(s^*) \geq 0$ for both types of the agent. Put differently, civil tests are so lenient that agents are acquitted with probability over a half regardless of types. After the guilty agent improves his testing technology, the principal considers the previous test threshold s^* insufficiently informative. To restore the test informativeness, the principal should design harder tests to elicit efforts from the innocent agent, and thereby improves the quality of learning. Formally,

Result 8. *When the guilty agent bears a lower effort cost, the equilibrium conviction threshold (conditional on its existence) is stricter in civil tests.*

3.5 Conclusion

In this paper, I develop a model to analyze the strategic interactions of testing. The key ingredient is to examine how a principal may set a private test standard when she anticipates an agent to influence test results with hidden efforts. I characterize conditions under which the principal employs a standard known as the "reasonable doubt" and convicts whenever the test result goes beyond it. The characterization suggests that, in spite of strategic concerns, the testing game is analytically equivalent to the principal's Bayesian decision problem.

I argue that we may not derive convincing comparative static results without specifying the test nature. As I demonstrate, a civil test differs from a criminal test with respect to whether a principal may use an extremely slack threshold to tell an innocent agent from a guilty one. Accordingly, in order to maintain the maximal possible leniency or deterrence, in different tests the principal may want to move her conviction threshold in the opposite directions owing to changes in the testing environment. This finding contrasts sharply with

conventional wisdom, which suggests that the principal unambiguously leans towards the lesser of two harms.

I believe there are a number of interesting avenues for future research – for instance, the power of commitment. My model assumes an implicit contract albeit without commitment between a principal and an agent. The no-commitment assumption is natural in the context of jury decision-making and electoral accountability. In other situations, it is more suitable to assume that the principal may have some commitment power. Commitment power is not always a blessing. For example, a legislature may draft media law to completely insulate their politicians from the disinformation war. But complete insulation also creates a moral hazard problem, because it imposes a constraint for media outlets to disclose information and thereby discipline politicians' behaviors. A full analysis of this trade-off awaits future research.

Declarations. The author is not aware of any memberships, funding, conflicts of interest that might affect the objectivity of this manuscript.

3.6 Appendix

3.6.1 Preliminaries

I state two useful facts with respect to the derivative of normal density ϕ. A function f is Lebesgue integrable if $\int_{\mathbb{R}} |f| < \infty$.

- $\phi^{(n)}$ is Lebesgue integrable.
- $\int_{\mathbb{R}} |\phi''(x)|dx \leq 1 + \frac{1}{2\sqrt{2}}$

Proof.

$$\int_{\mathbb{R}} |\phi''(s)|ds = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} |(s^2-1)e^{-s^2}|ds \leq \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} s^2 e^{-s^2} ds + 1 = \frac{1}{2\sqrt{2}} + 1 < \infty$$

□

3.6.2 Equilibrium characterization

First, we need to make sure that agents always have a well-defined objective.

Lemma 11 (Well-defined problem)**.** *The agent always admits a well-defined best response correspondence.*

In the proof, I suppress all type-subscripts.

Proof. G denotes the (measurable) set of all evidence realization $s = -h + \epsilon \in G$ at which conviction would be triggered. Hence, the probability of conviction $F(h)$ at evidence level h would be

$$\begin{aligned} F(h) &= \int 1\{s \in G\} d\Phi(\epsilon) = \int 1\{-h + \epsilon \in G\} d\Phi(\epsilon) \\ &= \int 1\{\epsilon \in G + h\} d\Phi(\epsilon) = \int_{G+h} \phi(\epsilon) d\epsilon \end{aligned}$$

$G + h$ is the right-shift of the set G by h. Rewrite $D = G + h$ and thus $F(h) = \int_D \phi(s)ds$. Now characterize the shape of F.

Claim 6. *For all* $h \in \mathbb{R}_+, |F'(h)| \leq 2, |F''(h)| \leq 1 + \frac{1}{2\sqrt{2}}$

Proof.

$$\begin{aligned} F'(h) &= \lim_{t\to 0} \int_D \frac{\phi(s+t) - \phi(s)}{t} ds \\ &= \int_D \lim_{t\to 0} \frac{\phi(s+t) - \phi(s)}{t} ds \\ &= \int_D \phi'(s) ds \\ &\leq \int_{\mathbb{R}} |\phi'(x)| dx = 2\int_{-\infty}^{0} \phi'(x) dx = 2\phi(0) \end{aligned}$$

The second line follows from Lebesgue dominant convergence theorem (using $1_D|\phi'(s+\xi)|$ with $\xi \in (0,t)$ as the dominating function. This is feasible due to mean value theorem). Similarly,

$$\begin{aligned} F''(h) &= \lim_{t\to 0} \int_D \frac{\phi'(s+t) - \phi'(s)}{t} ds \\ &= \int_D \lim_{t\to 0} \frac{\phi'(s+t) - \phi'(s)}{t} ds \\ &= \int_D \phi''(s) ds \\ &\leq \int_{\mathbb{R}} |\phi''(s)| ds \leq 1 + \frac{1}{2\sqrt{2}} \end{aligned}$$

□

With these results, agents' marginal gain in terms of favorable evidence-shift F' is uniformly bounded by $2\phi(0)$. On the other hand, their marginal cost to effort $C'(h)$ being

convex satisfies

$$C'(h) \geq C'(h(0)) + C''(h(0))(h - h(0)) = C''(0)(h - h(0))$$

As such, agents' choice is actually restricted to a compact set $[h(0), h(0) + \bar{h}]$, with the upper bound $\bar{h}$ given by $2\phi(0)/C''(0)$. This is because outside this compact set, the net marginal gain is always negative. Hence, the agents admit well-defined maximization problems. If we impose Assumption 1, then agent's problem is **globally concave**, thus admitting a unique maximizer. □

Principal's cutoff strategy

I start with a crucial observation: in any equilibrium, the innocent agent would appear more innocent than the gulty.

Lemma 12. *Fixing an arbitrary conviction strategy σ, it must be that the $h^*_\theta(\sigma)$ is uniquely selected and increasing in θ.*

Proof of Lemma 12. For $\theta \in \{I, G\}$

$$\begin{aligned} u_\theta(\sigma, h) =& Prob\{s \in G^c\} - C_\theta(h) \\ =& Prob\{\epsilon \in G^c + h\} - C_\theta(h) \end{aligned}$$

G^c is **independent** of θ due to principal's imperfect observability of type. By the SMP, u_θ has the increasing difference condition defined in Milgrom and Shannon [1994]. By Theorem 3 in Milgrom and Shannon [1994] it has the single crossing property. Furthermore, being a real-valued function $u_\theta(\sigma, h)$ is supermodular in h.

Apply the monotone selection theorem (Theorem 4') in Milgrom and Shannon [1994] to obtain $h^*_I \geq h^*_G$ as an elements of maximizer. By Theorem 3 in Edlin and Shannon [1998],

$h_I^* \neq h_G^*$. Combine two observations one obtains that $h_I^* > h_G^*$. □

Because lower s is always a better indicator of an innocent type, any equilibrium involving convicting at lower results and acquitting at higher results could be improved upon by swapping the conviction decision. Formally,

Lemma 13 (cutoff)**.** *Any equilibrium involves the principal using a cutoff strategy.*

I suppress the subscript of s_θ in the proof below.

Proof. Suppose not. Then there must be two disjoint intervals I, I' with $s < s'$ for all $s \in I, s' \in I'$ such that the principal convicts at s but acquit at s' . Since $s = -h + \epsilon$ and $h_I^* > h_G^*$, by Proposition 4 in Milgrom [1981] s' is a more "favorable" signal in the sense of being the criminal type. Let G denote the posterior distribution of type $\theta = C$ after the principal observes evidence, then the "favorableness" implies that $G(\cdot|s')$ dominates $G(\cdot|s)$ in the sense of strict first-order stochastic dominance (FOSD).

Now consider the new conviction strategy: fixing action after observing $s \notin I \cup I'$, but convict at s' and acquit at s. Under the new strategy, an additional $[G(C|s' \in I) - G(C|s \in I)]$ fraction of bad people are convicted correctly (and vise versa, the same fraction of good people are acquitted correctly). By FOSD, this quantity is strictly positive[6], thus strictly improving the principal's utility. □

This lemma establishes that we can without loss of generality restrict attention to equilibrium involving the principal's cutoff strategy. From now on, I write s^* as the principal's strategy $\mathcal{G} = \{s \leq s^*\}$. As such, I rewrite the payoffs in terms of the normal CDF Φ for a

6. To see it, $G(\cdot|s \in I)$ is the integral of an indicator function on the interval I against the posterior CDF of θ.

generic s:

$$u_\theta(s,h) = \Phi(s+h) - C_\theta(h) \quad \forall \theta \in \{I,G\}$$
$$u_p(s,h_I,h_G) = \Phi(s+h_I) - \alpha\Phi(s+h_G)$$

Equilibrium as correct inference

I start equilibrium characterization by looking at the first order conditions for each player, and analyze the possibility of an interior solution. It is easily checked that the FOC of three players are given as follows:

$$\text{Principal} \qquad \frac{\phi(s+h_I)}{\phi(s+h_c)} = \alpha \tag{3.1}$$
$$\text{Agent} \qquad \phi(s+h_\theta) = C'_\theta(h_\theta) \quad \theta \in \{I,G\} \tag{3.2}$$

Let us revisit the principal's problem. While it is *a priori* not clear whether her optimization problem is strictly globally concave, in fact the FOC is actually necessary and sufficient thanks to the MLRP.

Lemma 14. *If there is an s^* that solves Equation* (3.1)*, then it is the unique maximizer of $u_p(s,h_I,h_G)$ if only if $h_I > h_G$.*

Proof. ($\Leftarrow$): Under $h_I > h_G$, $\mathcal{L}(s,h_I,h_G) = \frac{\phi(s+h_I)}{\phi(s+h_G)}$ is strictly decreasing in s. For all $s < s^*$, $\mathcal{L}(s,h_I,h_G) > \alpha$ by MLRP so that u_p is increasing; for all $s > s^*$, u_p is decreasing. Hence, at s^* P maximizes her utility u_p. In other words, the s^* solving first order condition is $P's$ best response to (h_I,h_G).

($\Rightarrow$) Suppose s^* is the maximizer that solves FOC but $h_I < h_G$, then u_p admits a global minimum at s^* using the similar argument. If $h_I = h_G$, then $\mathcal{L} = 1 > \alpha$. Hence, no equilibrium is feasible. □

Now we may verify the the conjectured equivalence between a testing game and a (modified) Bayesian inference problem.

Proof of Proposition 3. Note that $h_I^* > h_G^*$ by Lemma 12. The result follows from the definition of equilibrium, and sufficiency of FOC established in Lemma 14. □

The proposition allows us to determine the existence property of an equilibrium by studying the shape of $LR(s)$, where

$$LR(s) = \frac{\phi(s + h_I^*(s))}{\phi(s + h_G^*(s))}$$

The LR differs from $\mathcal{L}$ in that $h_\theta^*(s)$ best responds to s. Hence, LR loses the MLRP.

Analyzing test informativeness

Non-monotone best response. I restate Lemma 9 as the following lemma:

Lemma 15. *The agent's best response to s, $h_\theta^*(s)$, is not monotone. In particular, there exists an $\hat{s}_\theta := -(C_\theta')^{-1}(\phi(0))$, such that $h_\theta^*(s)$ is increasing in s if $s \leq \hat{s}_\theta$, and decreasing if $s > \hat{s}_\theta$. Finally, $\hat{s}_I < \hat{s}_G$.*

To gain some intuition, consider the incentive of an agent when s decreases (harder test). On the one hand, she/he may either increase effort to counteract s, or decrease effort *further* to exploit the "economy" of saving effort cost. Whichever effect dominates depends on the test difficulty.

Proof of Lemma 15. I suppress type in this section.
Consider agents' best response functions $\phi(s + h) = C'(h)$ and SOC $\phi' < C''$ where $\phi' = \phi'(s + h)$. One can check the derivative of h with respect to s

$$h'(s) = \frac{\phi'}{C'' - \phi'}$$

It is clear that $\phi'(x) < (>)0$ for all $x > (<)0$. Therefore one concludes

$$h'(s) = \begin{cases} -\frac{|\phi'|}{C''+|\phi'|} \in (-1, 0] \text{ for } s + h \geq 0 \\ \frac{\phi'}{C''-\phi'} > 0 \text{ for } s + h < 0 \end{cases} \tag{3.3}$$

If we define $\hat{s} = -(C'_\theta)^{-1}(\phi(0))$, then h is increasing in s if $s \leq \hat{s}$ and decreasing otherwise. Using the SMP, one concludes that $\hat{s}_I < \hat{s}_G$. □

Intuitively, harder tests always lead to more conviction:

Corollary 2. *$s + h^*_\theta(s)$ is increasing in s. Therefore, the conviction rate $1 - \Phi(s + h^*_\theta(s))$ is always decreasing in s.*

Proof. The derivative of $s + h(s), h'(s) + 1$, is positive by (3.3). □

Interior behavior. We can establish its "interior behavior" by exploiting the single-peakedness of ϕ.

Lemma 16 (Interior behavior)**.** *There exists $\underline{s} < \bar{s}$ such that $LR(\underline{s}) > 1, LR(\bar{s}) < 1$. Furthermore, $LR(s)$ is decreasing for $s \in (\underline{s}, \bar{s})$*

Proof of Lemma 16. Consider equation (3.2). Take $\underline{s} = \hat{s}_I, \bar{s} = \hat{s}_c$. At $\hat{s}_\theta$, type θ will best respond by choosing $h^*_\theta = -\hat{s}_\theta$, resulting a density $\phi = \phi(0)$ which is maximized by single-peakedness. $\bar{s}$ corresponds to $\hat{s}_G$ with the induced $LR(\bar{s}) < 1$, and $\underline{s}$ corresponds to $\hat{s}_I$ with the induced $LR(\underline{s}) > 1$.

Turn to the second part. By Corollary 2, $s+h^*(s)$ is increasing in s. By definition of $\underline{s}, \bar{s}$, $\underline{s}+h_I(\underline{s}) = 0$ and $\bar{s}+h_G(\bar{s}) = 0$. Hence for any $s \in (\underline{s}, \bar{s})$, $s+h_I(s) > 0$ but $s+h_G(s) < 0$. □

Asymptotic behavior. We say that the innocent and the guilty agents are *asymptotically indistinguishable* if $LR(s) \uparrow 1^-$ for $s \uparrow \infty$. The definition is tested against the limiting

behavior of LR[7]. It basically checks whether an extremely easy test is *uninformative*.

Looking complicated, the condition has a surprising simple characterization:

Lemma 17 (Asymptotic behavior)**.** *Agents are asymptotically indistinguishable if and only if $h_0 = 0$. In other words, the innocent has no initial advantage. Formally,*

- *If $h_0 = 0$, then $LR(s) \uparrow 1^-$ for $s \uparrow \infty$. It also admits a minimum at some $\exists s_0 \geq \bar{s}$.*
- *If $h_0 > 0$, then $LR(s) \in (0, 1)$, and $LR(s) \downarrow 0^+$ for $s \uparrow \infty$.*

Proof. Case I: $h_0 = 0$. We can without loss assume $h_\theta(0) = 0$.

$h_I^* > h_c^*$ is immediate by Lemma 12. Note that for $s > \bar{s}$,

$$1 > LR(s) = \frac{\phi(h_I^* + s)}{\phi(h_G^* + s)} \geq \frac{\phi(h_I^* + s)}{\phi(s)} = \frac{\exp[-\frac{(s+h_I^*)^2}{2}]}{\exp[-\frac{s^2}{2}]} = \exp[-h_I^* s] \exp[-(h_I^*)^2/2]$$

Now let $s \to +\infty$. Note that $h_I^* \to 0$ as $s \to +\infty$ from Equation (3.2). It suffices to show that $h_I^* s \to 0$ as $s \to +\infty$. In other words, h_I^* decays faster than $\frac{1}{s}$.

Now bound the size of sh_I^*. The first order condition for the agent is $\phi(s + h) = C'_\theta(h)$. By convexity of C'_θ, $C'_\theta(h) \geq C'_\theta(0) + C''_\theta(0)h$, where $C''(0) > 0$. As such, for $s >> \bar{s}$

$$h_I^* \leq \frac{1}{C''_\theta(0)} \phi(s + h_I^*) \leq \frac{1}{C''_\theta(0)} \phi(s)$$

Hence

$$\lim_{s\to\infty} h_I^* s \leq \lim_{s\to\infty} s \frac{1}{\sqrt{2\pi}} \frac{1}{C''_\theta(0)} \exp[-\frac{s^2}{2}] = \frac{1}{\sqrt{2\pi}} \frac{1}{C''_\theta(0)} \lim_{s\to\infty} \frac{s}{\exp[s^2/2]} \to 0$$

Next, show $\exists s_0 \geq \bar{s}$ such that $LR(s)$ attains its minimum. Pick $\epsilon < 1 - LR(\bar{s})$. By indistinguishability condition, $\exists s^*$ such that $LR(s) > 1 - \epsilon$ for all $s \geq s^*$. By construction,

7. We focus on the limit $s \uparrow +\infty$. The other limiting property can be defined similarly, but it is not relevant for the equilibrium characterization of $\alpha \in (0, 1)$ due to the single-peakedness of normal pdf. The likelihood ratio always exceeds 1 there.

$LR(s^*) > LR(\bar{s})$. Moreover, note that LR is decreasing for $s \in (\underline{s}, \bar{s}]$. Hence, $LR(s)$ must attain its minimum $s_0 \in [\bar{s}, s^*]$.

Case II: $h_0 > 0$.

Consider the case where $s \uparrow \infty$. Note that for any $h_0 := h_I(0) > 0$ we can pick s sufficiently large such that $h_G^* \leq h_0/2$

$$LR(s) = \frac{\phi(h_I^* + s)}{\phi(h_G^* + s)} \leq \frac{\phi(h_0 + s)}{\phi(h_0/2 + s)}$$

Since $h_0 > 0$ is a constant, RHS clearly approaches 0 as $s \uparrow \infty$. The case for $s \downarrow -\infty$ is similar. □

Due to the symmetry of $LR(s)$, Lemma 16 and Lemma 17 completely characterize the range of $LR(s)$. As such, we may prove Proposition 4.

Proof of Proposition 4. Civil test ($h_0 = 0$). Note that Lemma 12 guarantees $h_I^* > h_c^*$ for any choice of $s > 0$. By Lemma 17 indistinguishability condition is met. By intermediate value theorem (IVT), any $\alpha \in [LR(\underline{s}), 1)$ would crosses $LR(s)$. *Criminal test ($h_0 > 0$).* By Lemma 17, any $\alpha \in (0, 1)$ will definitely cross $LR(s)$. □

Lemma 18. *In civil tests, all agents in equilibrium are acquitted with probability at least one half.*

Proof. Consider nontrivial equilibria only. Recall that $LR(s)$ approaches 1 for s close to s_1 or ∞. Since $LR(s)$ is decreasing for $s \in (s_1, \bar{s})$ and s_0 achieves the minimum value for all $s \in (s_1, \infty)$, it must be that $s_0 \geq \bar{s}$. At $s = \bar{s}$, by definition the guilty agent is acquitted with probability 1/2. For $s \geq \bar{s}$, applying Corollary 2 it must be that the guilty agent is acquitted with a lower probability than 1/2. □

3.6.3 Comparative Statics

Blackstone Ratio

Restate Implication 7: *The (selected) equilibrium s and conviction probability is*

- *decreasing in α in criminal tests;*
- *increasing in α in civil tests.*

Proof. I prove the case of civil tests. The other is analogous since we have selected on the extreme equilibria. In civil tests, $LR(s) \uparrow 1^-$ as $s \uparrow \infty$.

Let (s, h) be the largest equilibrium and denote the new equilibrium (s', h') for some $1 > \alpha' > \alpha$. This implies that $LR(s') = \alpha'$ and $LR(s) = \alpha$. Assume towards contradiction that $s' < s$. By IVT, there exists $s'' > s$ such that $LR(s'') = \alpha'$, contradicting that s' is the largest with respect to α'. □

Testing technology

Restate Implication 8: *When the guilty agent has a lower effort cost, the equilibrium conviction threshold is stricter only in civil tests. .*

Proof. Use $h^*_G(s), h'^*_G(s)$ to denote guilty agent's old/new best response, and s^*, s'^* old/new equilibrium conviction threshold. Note first that $h'_G > h_G$ by the Theorem 3 in Edlin and Shannon [1998].

Now I focus on the civil test comparative statics. Under $1 > \alpha > LR(\underline{s})$, and because we have selected the largest equilibrium s^*, it must be that $s^* + h^*(s^*) > 0$ and so $s^* + h'^*(s^*) > 0$. $\frac{\phi(s^*+h^*_I(s^*))}{\phi(s^*+h^*_G(s^*))} = \alpha < \frac{\phi(s^*+h^*_I(s^*))}{\phi(s^*+h'^*_G(s^*))}$ because $h'_G(s) > h_G(s)$. Suppose towards contradiction that $s'^* > s^*$. However, this implies the existence of an $s'' \in (s^*, \infty)$ such that $\frac{\phi(s''+h^*_I(s''))}{\phi(s''+h'^*_G(s''))} = \alpha$, contradicting the s'^* being the largest extreme equilibrium. □

3.6.4 Robustness of results

Throughout the paper, I assume that the noise ϵ is distributed according to standard Gaussian. One may wonder how results are robust to alternative noise distributions. I first discuss the role of Gaussian noise, followed by generalizing main results to a broader class of distributions.

In the proof, Gaussian noise plays three roles.

- The pdf $\phi(x)$ belongs to the Schwartz space[8], ensuring that ϕ', ϕ'' are Lebesgue integrable. This property helps bound agents' marginal gains in terms of changing conviction probability, making their optimization problem well-defined. With stronger assumptions on the shape of cost function, agents' problem could even be made concave.

- Being in the Schwartz space, ϕ decays very fast at infinity. Since agent's marginal cost function dominates a linear one, it can be shown that at $s \uparrow \infty$, agents of two types put so little effort that they are indistinguishable when $h_0 = 0$.

- The pdf $\phi(x)$ has MLRP. Hence the first order condition pins down principal's best response.

Acknowledging this, we can generalize noise distributions to the following class:

Theorem 4. *Suppose ϵ admits absolute continuous cumulative density function F with pdf f satisfying the following properties:*

1. *$f \in S(\mathbb{R})$, where $S(\mathbb{R})$ is the Schwartz space.*

2. *f is atomless, single-peaked, symmetric around 0 with unbounded support, and has MLRP.*

8. See [Folland, 2007, 237] for integration properties of Schwartz class.

3. $\frac{f'(x)}{f(x)} = P(x)$*, where* $P(x)$ *is a Laurent polynomial with positive degrees*[9]*.*

Replacing the Gaussian assumption with ϵ *distributed as such, then all conclusions hold.*

These requirements encompass many distributions with unbounded support that we are familiar with, including Gaussian, Laplace distribution, modified Gamma and so on. Let us prove the robustness of results according to different distributions:

Proof of Theorem 4. The agent has a well-defined problem and a unique selection. Monotonicity arguments follow, which establishes the optimality of BARD.

No initial advantage. Single-peakedness of f ensures Lemma 16. Now check the indistinguishability condition, or whether $LR(s) \to 1$ as $s \uparrow \infty$. Denote $L(x) = \log f(x)$. Since for s sufficiently large under BARD, $h^* \to 0$. Therefore $s + h^* > 0$ and thus f, L are both decreasing when $s \uparrow \infty$. It suffices to show that $L(s + h_I^*) - L(s + h_G^*) \to 0$ as $s \to 0$.

$$0 \geq L(s + h_I^*) - L(s + h_G^*) \geq L(s + h_G^*) - L(s) = L'(s + \xi(s))h_I^*, \quad \xi(s) \in (0, h_I^*)$$

By Equation (3.2), $0 \leq h_I^* \leq \frac{1}{C_I''(0)} f(s)$. Hence,

$$\begin{aligned} L(s + h_I^*) - L(s + h_G^*) &\geq \frac{1}{C_I''(0)} L'(s + \xi) f(s) = \frac{1}{C_I''(0)} f'(s + \xi) \frac{f(s)}{f(s + \xi)} \\ &= \frac{1}{C_I''(0)} [P(s + \xi)] f(s) \end{aligned}$$

First note that we can without loss assume that P is indeed a polynomial. This is because any term of $(s + \xi)$ with negative order is of the form $[\frac{1}{s+\xi}]^n$ which is bounded by some constant for s sufficient large.

Next, recall in the proof of Lemma 12, we know that for any $\bar{h}$ there exists $\tilde{s}$ such that $s \geq \tilde{s}$ implies $h_I^* < \bar{h}$. This observation helps bound $s + \xi$. Now, I show that there

9. Laurent polynomials are polynomials potentially with negative degrees. The "positive" part in this statement is almost superfluous, for otherwise the density is not integrable.

exist polynomials $\bar{P}, \underline{P}$ such that $\underline{P}(s) \leq P(s+\xi) \leq \bar{P}(s)$ for all $s > 0$. Write $P(s+\xi) = a_0 + a_1(s+\xi) + ... + a_n(s+\xi)^n$. Note that since $s > 0, s+\xi > 0$, we can define $\bar{P}(s) = |a_0| + |a_1|(s+\bar{h}) + ... + |a_n|(s+\bar{h})^n$ and $\underline{P} = -\bar{P}$. Both are polynomials. Since $f \in S(\mathbb{R}), \lim_{s\to\infty} \bar{P}(s)f(s) = \lim_{s\to\infty} \underline{P}(s)f(s) = 0$, Hence, $\lim_{s\to\infty} P(s+\xi)f(s) = 0$. This proves indistinguishability condition.

Together with Lemma 16, we have two crossing points of α with respect to LR function. By MLRP, the principal is best responding at these points. As such, Proposition 4 applies. Comparative static analysis follows.

Initial advantage In the same vein, it suffices to check if $L(s+h_I^*) - L(s+h_G^*) \to -\infty$ as $s \uparrow \infty$. Using a similar argument, $L(s+h_I^*) - L(s+h_G^*) = P(s+\xi)$ for some $\xi(s) \in (0, h_I^*)$. Since P has positive degree, $P(x+\xi)$ is not bounded as $s \uparrow \infty$. Its limit can only be $-\infty$, since $L(x)$ is a decreasing function for $x \geq 0$. The rest of arguments follows. □

REFERENCES

Avidit Acharya and Edoardo Grillo. A Behavioral Foundation for Audience Costs. *Quarterly Journal of Political Science*, 14(2):159–190, 2019. ISSN 1554-0626.

Scott Ashworth. Electoral Accountability: Recent Theoretical and Empirical Work. *Annual Review of Political Science*, 15(1):183–201, 2012.

Scott Ashworth and Ethan Bueno de Mesquita. Is Voter Competence Good for Voters?: Information, Rationality, and Democratic Performance. *American Political Science Review*, 108(3):565–587, 2014.

Scott Ashworth and Kristopher Ramsay. Should Audiences Cost? Optimal Domestic Constraints in International Crises. *mimeo*, 2017. URL `http://home.uchicago.edu/~sashwort/audience.pdf`.

Scott Ashworth, Ethan Bueno de Mesquita, and Amanda Friedenberg. Accountability and Information in Elections. *American Economic Journal: Microeconomics*, 9(2):95–138, May 2017a.

Scott Ashworth, Ethan Bueno de Mesquita, and Amanda Friedenberg. Learning about Voter Rationality. *American Journal of Political Science*, 62(1):37–54, 2017b.

Jeffrey Banks. Equilibrium Behavior in Crisis Bargaining Games. *American Journal of Political Science*, 34(3):599–614, 1990.

Jeffrey S. Banks and Joel Sobel. Equilibrium Selection in Signaling Games. *Econometrica*, 55(3):647–661, 1987.

Robert Barro. The Control of Politicians: An Economic Model. *Public Choice*, 14:19–42, 1973.

Gary Becker. Crime and Punishment: An Economic Approach. *Journal of Political Economy*, 76(2):169–217, 1968.

Timothy Besley. *Principled Agents?: The Political Economy of Good Government.* Oxford University Press, 2006.

William Blackstone. *Commentaries On the Laws of England.* Boston :Beacon Press, 1962.

Geofferey Blainey. *The Causes of War.* Free Press, 3rd, 1988.

Tilman Borgers, Daniel Krahmer, and Roland Strausz. *An Introduction to the Theory of Mechanism Design.* New York : Oxford University Press, 2015.

Bruce Bueno de Mesquita, James Morrow, Randolph Siverson, and Alastair Smith. An Institutional Explanation of the Democratic Peace. *The American Political Science Review*, 93(4):791–807, 1999.

Ester Camiña and Nicolás Porteiro. The Role of Mediation in Peacemaking and Peacekeeping Negotiations. *European Economic Review*, 53(1):73 – 92, 2009.

Brandice Canes-Wrone, Michael C. Herron, and Kenneth W. Shotts. Leadership and Pandering: A Theory of Executive Policymaking. *American Journal of Political Science*, 45 (3):532–550, 2001.

Marquis de. Condorcet. *Essai sur l'application de l'analyse a la probabilite des decisions rendues a la probabilite des voix. Paris: De l'imprimerie royale, translated in 1976 to " Essay on the Application of Mathematics to the Theory of Decision-Making." in Condorcet: Selected Writings, ed. Keith M. Baker.* Indianapolis, IN: Bobbs-Merrill., 1785.

Peter Coughlan. In Defense of Unanimous Jury Verdicts: Mistrials, Communication, and Strategic Voting. *The American Political Science Review*, 94(2):375–393, 2000.

Jacques Crémer. Arm's Length Relationships. *The Quarterly Journal of Economics*, 110(2): 275–295, 1995.

Mathias Dewatripont, Ian Jewitt, and Jean Tirole. The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies. *The Review of Economic Studies*, 66(1):199–217, 1999.

Aaron Edlin and Chris Shannon. Strict Monotonicity in Comparative Statics. *Journal of Economic Theory*, 81(1):201 – 219, 1998.

Georgy Egorov. Single-Issue Campaigns and Multidimensional Politics. *mimeo*, 2015. URL `https://www.kellogg.northwestern.edu/faculty/egorov/ftp/Single-Issue%20Campaigns.pdf`.

Katja Favretto. Should Peacemakers Take Sides? Major Power Mediation, Coercion, and Bias. *The American Political Science Review*, 103(2):248–263, 2009.

James Fearon. Domestic Political Audiences and the Escalation of International Disputes. *The American Political Science Review*, 88(3):577–592, 1994.

James Fearon. Rationalist Explanations for War. *International Organization*, 49(3):379–414, 1995.

James Fearon. Electoral accountability and the control of politicians: Selecting good types versus sanctioning poor performance. In Adam Przeworski, Susan C. Stokes, and BernardEditors Manin, editors, *Democracy, Accountability, and Representation*, Cambridge Studies in the Theory of Democracy, page 55–97. Cambridge University Press, 1999. doi: 10.1017/CBO9781139175104.003.

Timothy Feddersen and Wolfgang Pesendorfer. Convicting the Innocent: The Inferiority of Unanimous Jury Verdicts under Strategic Voting. *The American Political Science Review*, 92(1):23–35, 1998.

John Ferejohn. Incumbent Performance and Electoral Control. *Public Choice*, 50(1/3):5–25, 1986.

Mark Fey and Kristopher Ramsay. Mechanism Design Goes to War: Peaceful Outcomes with Interdependent and Correlated Types. *Review of Economic Design*, 2009.

Mark Fey and Kristopher Ramsay. When Is Shuttle Diplomacy Worth the Commute? Information Sharing through Mediation. *World Politics*, 62(4):529–560, 2010.

Mark Fey and Kristopher Ramsay. Uncertainty and Incentives in Crisis Bargaining: Game-Free Analysis of International Conflict. *American Journal of Political Science*, 55(1): 149–169, 2011.

Gerald Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley; 2nd edition, 2007.

Justin Fox. Government Transparency and Policymaking. *Public Choice*, 131(1/2):23–44, 2007.

Justin Fox and Stuart V. Jordan. Delegation and Accountability. *The Journal of Politics*, 73(3):831–844, 2011.

Justin Fox and Kenneth W. Shotts. Delegates or Trustees? A Theory of Political Accountability. *The Journal of Politics*, 71(4):1225–1237, 2009.

Justin Fox and Richard Van Weelden. Costly transparency. *Journal of Public Economics*, 96(1):142–150, 2012. ISSN 0047-2727.

Drew Fudenberg and Jean Tirole. *Game Theory*. MIT Press, Cambridge, MA, 1991.

Dino Gerardi and Leeat Yariv. Information Acquisition in Committees. *Games and Economic Behavior*, 62(2):436 – 459, 2008.

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

Devin T Hagerty. *The Consequences of Nuclear Proliferation*. MIT Press, 1998.

Alexander V. Hirsch. Experimentation and Persuasion in Political Organizations. *American Political Science Review*, 110(1):68–84, 2016.

Bengt Holmström. Managerial Incentive Problems: A Dynamic Perspective. *The Review of Economic Studies*, 66(1):169–182, 1999.

Bengt Hölmstrom. Moral Hazard and Observability. *The Bell Journal of Economics*, 10(1): 74–91, 1979.

Johannes Hörner, Massimo Morelli, and Francesco Squintani. Mediation and Peace. *The Review of Economic Studies*, 82(4 (293)):1483–1501, 2015.

Matthew Jackson and Massimo Morelli. Political Bias and War. *American Economic Review*, 97(4):1353–1373, 2007.

Junyan Jiang and Jeremy Wallace. Informal Institutions and Authoritarian Information Systems: Theory and Evidence from China. *mimeo*, 2017. URL `https://dx.doi.org/10.2139/ssrn.2992165`.

Louis Kaplow. On the Optimal Burden of Proof. *Journal of Political Economy*, 119(6): 1104–1140, 2011.

Louis Kaplow. Optimal Multistage Adjudication. *The Journal of Law, Economics, and Organization*, 33(4):613–652, 2017.

Louis Kriesberg. Mediation and the Transformation of the Israeli-Palestinian Conflict. *Journal of Peace Research*, 38(3):373–392, 2001.

Andrew Kydd. When Can Mediators Build Trust? *The American Political Science Review*, 100(3):449–462, 2006.

Henrik Lando. Does Wrongful Conviction Lower Deterrence? *The Journal of Legal Studies*, 35(2):327–337, 2006.

Jeffrey Lax. Political Constraints on Legal Doctrine: How Hierarchy Shapes the Law. *The Journal of Politics*, 74(3):765–781, 2012.

Luis Martinez. How Much Should We Trust the Dictator's GDP Growth Estimates? *mimeo*, 2019. URL `http://dx.doi.org/10.2139/ssrn.3093296`.

Eric Maskin and Jean Tirole. The Politician and the Judge: Accountability in Government. *American Economic Review*, 94(4):1034–1054, September 2004.

Paul Milgrom. Good News and Bad News: Representation Theorems and Applications. *The Bell Journal of Economics*, 12(2):380–391, 1981.

Paul Milgrom and Ilya Segal. Envelope Theorems for Arbitrary Choice Sets. *Econometrica*, 70(2):583–601, 2002.

Paul Milgrom and Chris Shannon. Monotone Comparative Statics. *Econometrica*, 62(1): 157–180, 1994.

Dilip Mookherjee and L. Png. Marginal Deterrence in Enforcement of Law. *Journal of Political Economy*, 102(5):1039–1066, 1994.

Stephen Morris. Political Correctness. *The Journal of Political Economy*, 109(2):231–265, 2001.

Roger Myerson. Incentive Compatibility and the Bargaining Problem. *Econometrica*, 47(1): 61–73, 1979.

Roger Myerson. Optimal Auction Design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

Roger Myerson. Two-Person Bargaining Problems with Incomplete Information. *Econometrica*, 52(2):461–487, 1984.

Roger Myerson and Mark Satterthwaite. Efficient Mechanisms for Bilateral Trading. *Journal of Economic Theory*, 29(2):265 – 281, 1983.

Nicola Persico. Committee Design with Endogenous Information. *The Review of Economic Studies*, 71(1):165–191, 2004.

Mattias Polborn and David Yi. Informative Positive and Negative Campaigning. *Quarterly Journal of Political Science*, 1(4):351–371, 2006.

Richard Posner. An Economic Approach to the Law of Evidence. *Stanford Law Review*, 51: 1477–1546, 1999.

Andrea Prat. The Wrong Kind of Transparency. *American Economic Review*, 95(3):862–877, June 2005.

Kristopher Ramsay. Information, Uncertainty, and War. *Annual Review of Political Science*, 20(1):505–527, 2017.

John G. Riley. Informational Equilibrium. *Econometrica*, 47(2):331–359, 1979.

Kenneth Schultz. Looking for Audience Costs. *The Journal of Conflict Resolution*, 45(1): 32–60, 2001.

Stergios Skaperdas. Contest Success Functions. *Economic Theory*, 7(2):283–290, 1996.

Michael Spence. Job Market Signaling. *The Quarterly Journal of Economics*, 87(3):355–374, 1973.

George Stigler. The Optimum Enforcement of Laws. *Journal of Political Economy*, 78(3): 526–36, 1970.

Ahmer Tarar and Bahar Leventoglu. Limited Audience Costs in International Crises. *The Journal of Conflict Resolution*, 57(6):1065–1089, 2013.

Michael Ting. Politics and Administration. *American Journal of Political Science*, 61(2): 305–319, 2017.

Kenneth Waltz. *Man, the State, and War: A Theoretical Analysis*. Columbia University Press, 2001.