Essays in Political Economy

by

Eitan Sapiro-Gheiler

B.A. Economics, Princeton University, 2019

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN ECONOMICS

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2025

Authored by:	Eitan Sapiro-Gheiler Department of Economics May 15, 2025
Certified by:	Alexander Wolitzky Professor of Economics, Thesis Supervisor
Accepted by:	Isaiah Andrews Charles E. and Susan T. Harris Professor of Economics Chairman, Departmental Committee on Graduate Studies

THESIS COMMITTEE

Thesis Supervisor

Alexander Wolitzky
Professor of Economics
Department of Economics

Committee Members

Benjamin Olken
Jane Berkowitz Carlton and Dennis William Carlton Professor of Microeconomics
Department of Economics

Stephen Morris
Peter A. Diamond Professor of Economics
Department of Economics

Essays in Political Economy

by

Eitan Sapiro-Gheiler

Submitted to the Department of Economics
on May 15, 2025 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN ECONOMICS

ABSTRACT

In this thesis, I describe three approaches to political communication and decision-making. Chapter 1, "Persuasion with Ambiguous Receiver Preferences," studies an informed Sender who knows only the average threshold belief needed to persuade a Receiver and wishes to safeguard against unfavorable distributions of individual preferences. Chapter 2, "Discovery through Trial Balloons," examines how correlation between different projects affects information disclosure by a principal who designs a bundle of projects that an agent can then choose to approve. Chapter 3, "Strategic Opinion-Writing on Appellate Courts," describes how and why the partisan composition of quasi-random panels of judges on the U.S. Federal Courts of Appeals affects consensus-building. I describe each chapter in more detail below.

The first chapter, "Persuasion with Ambiguous Receiver Preferences," describes a Bayesian persuasion problem where Receiver has a private belief cutoff for Sender's preferred action and Sender has maxmin preferences over all Receiver type distributions with known mean and bounds. This problem can be represented as a zero-sum game where Sender chooses a mean-preserving contraction of the prior over states and adversarial Nature chooses a Receiver type distribution. I formalize the connection between maxmin persuasion and similar games used to model political spending, all-pay auctions, and competitive persuasion. In both a standard binary-state setting and a new continuous-state setting, Sender optimally linearizes the prior distribution over states to create a distribution of posterior means that is uniform on a known interval with an atom at the lower bound of its support.

The second chapter, "Discovery through Trial Balloons," presents a model of a principal and an agent who face symmetric uncertainty about the agent's value for two correlated projects. The principal chooses which project values to publicly discover and makes a proposal to the agent, who accepts if and only if the expected sum of values is positive. I characterize optimal discovery for various principal preferences: maximizing the probability of the grand bundle, of having at least one project approved, and of a weighted combination of projects. My results show when discovering ex-ante disfavored projects may be optimal; these conclusions rationalize the inclusion of controversial policies in omnibus bills and the presence of moonshot projects in organizations.

The third chapter, "Strategic Opinion-Writing on Appellate Courts," studies consensus and decision-making by powerful judges on the U.S. Federal Courts of Appeals. Using quasi-random three-judge panels on these courts from 1970–2013 I document a novel pattern in dissenting opinions: compared to party-unanimous panels, party-mixed panels cause all judges to dissent more often, and at equal rates. This result is incompatible with classical

models of judicial politics and is unique to partisanship. To explain my results, I introduce a theoretical framework where judges' favored coalitions are more homogeneous along both partisan and non-partisan dimensions. Using judge metadata, I find suggestive evidence for the model's result that polarization increases dissents by judges of panel-minority law school or gender. With state-of-the-art machine learning tools from natural language processing, I generalize beyond dissents, showing that those same features drive differences in opinion text even when rulings are unanimous.

Thesis supervisor: Alexander Wolitzky
Title: Professor of Economics

Acknowledgments

"Stars, wind, polarised light, magnetic cues, the distant rubble of clouds a hundred miles out, clear cold air, and below them the hush of a world tilting towards sleep or waking towards dawn."

—Helen Macdonald, *Vesper Flights* [62].

Contents

List of Figures

List of Tables

Chapter 1

Persuasion with Ambiguous Receiver Preferences

I describe a Bayesian persuasion problem where Receiver has a private type representing a cutoff for choosing Sender's preferred action, and Sender has maxmin preferences over all Receiver type distributions with known mean and bounds. This problem can be represented as a zero-sum game where Sender chooses a distribution of posterior mean beliefs that is a mean-preserving contraction of the prior over states, and an adversarial Nature chooses a Receiver type distribution with the known mean; the player with the higher realization from their chosen distribution wins. I formalize the connection between maxmin persuasion and similar games used to model political spending, all-pay auctions, and competitive persuasion. In both a standard binary-state setting and a new continuous-state setting, Sender optimally linearizes the prior distribution over states to create a distribution of posterior means that is uniform on a known interval with an atom at the lower bound of its support.

JEL Classification: D81, D82, D83
Keywords: Bayesian persuasion, maxmin utility, mean-preserving contraction

Chapter Acknowledgments: I thank Drew Fudenberg, Stephen Morris, Frank Schillbach, Dmitry Taubinsky, participants in MIT Theory Lunch, participants in MIT 14.192, and especially Alexander Wolitzky for helpful discussions and comments. I also thank two anonymous referees for their feedback and suggestions. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1745302.

1.1 Introduction

Consider a politician who is deciding how to disclose information about the cost-effectiveness of a new welfare program, but does not know how much spending voters will support. All voters have the same prior beliefs, but some will only approve if they expect the program to provide a high level of benefits per dollar spent, while others are willing to support even a moderately inefficient government outlay. Rather than imposing a prior distribution over preferences, the politician wishes to be robust to the worst-case distribution she may face given a known threshold for the average voter. In this setting, what disclosure rule maximizes the share of voters who approve of the welfare program after taking into account the politician's message? How do the optimal rule and the politician's utility differ from the case where the politician faces a known distribution of citizen preferences?

I address and generalize those questions through a model of Bayesian persuasion [48], where a Sender commits to a message distribution in each state of the world and a Receiver uses Bayesian updating to form a posterior belief about the state based on the message structure. To represent Receiver's preferences, I use private types denoting the cutoff above which Receiver chooses Sender's preferred action. Sender knows the mean and support of Receiver types, and has maxmin preferences [40] over all Receiver type distributions satisfying those constraints. Regardless of the true state of the world, Sender maximizes the probability of inducing the favorable action. This model captures situations where all Receiver types process information in the same way, but may have different preferences over outcomes. In addition to the political spending example described above, a model of this style also applies to a variety of other situations, such as disclosing information about product quality (if potential customers share a prior belief about quality, but may be more or less picky about when they buy) or screening job candidates (if all firms have a common prior about candidate quality and see the same resumé, but have different thresholds for hiring).

This persuasion model can be reinterpreted as a zero-sum game between Sender and an adversarial Nature. Following the Bayesian persuasion literature, I can allow Sender to directly choose any distribution of posterior mean beliefs about the state that is a mean-preserving contraction of the prior. Then, Nature chooses a Receiver type distribution with the appropriate mean and domain; this choice is equivalent to choosing a mean-preserving contraction of a Receiver type distribution with support $\{0, 1\}$. The player with the higher realization from their chosen distribution wins the game. Such *mean-preserving contraction games* (henceforth MPC games), albeit with simultaneous moves, have been studied in prior literature outside of the persuasion context (for example by Myerson 64), as well as being used to represent competition between many Senders persuading a single Receiver (as in Boleslavsky and Cotton 14). Many of those works emphasize the role of uniform distributions, which induce indifference among many possible strategies for the opposing players. Adapting these results to my setting, I show that in a binary-state setting where the probability of the high state is weakly less than $1/2$, Sender's unique optimal posterior distribution places an atom at 0 and is uniform on an interval $[0, c]$ for $c \leq 1$. In doing so, I formalize the connection between maxmin persuasion and MPC games and show that the sequential timing of the maxmin persuasion game does not affect Sender's optimal distribution but the tie-breaking rule sometimes does. I also use a geometric approach based on the concavification argument

of Kamenica and Gentzkow [48] to show that for any finite number of states of the world, or when the state is continuous and unimodal, a similar distribution—uniform on $[a, b] \subset (0, 1)$ with an atom at a—is one of many optimal distributions for Sender. The continuous-state setting is a novel specification of both the MPC game and the maxmin persuasion problem.

1.2 Related Literature

This work builds on the Bayesian persuasion problem of Kamenica and Gentzkow [48], and adopts a similar approach to existing work in robust mechanism design. In addition, my model resembles a class of games I call MPC games, which include a continuous version of the Colonel Blotto game as well as competitive Bayesian persuasion by multiple Senders. I discuss the first two topics here and postpone discussion of the third to Section 1.3.3, after presenting the formal model.

In the baseline Bayesian persuasion model of Kamenica and Gentzkow [48], Receiver has no private information. Subsequent literature in this area is surveyed in detail by Kamenica [47] and Bergemann and Morris [11], so I focus on the two works most directly related to the model I propose, Kolotilin et al. [53] and Hu and Weng [44].[1] The former has an interval state space, Receiver types that enter payoffs linearly, and a binary action, as in my model; however, it endows Sender with a prior distribution over Receiver types. If that prior distribution is log-concave, then the optimal distribution for Sender can be generated by upper censorship; the resulting distribution of posterior means is essentially a truncated version of the prior where states in some interval $[\alpha, 1]$ are replaced with an atom at $\beta \in (\alpha, 1)$. In the continuous-state version of my model, linearizing the prior rather than censoring high states helps Sender avoid facing a tailored Receiver type distribution in response. To make sure this strategy respects Bayes-plausibility, Sender may use a truncated uniform distribution with interior support.

The model of Hu and Weng [44] is most similar to the one considered here: it is a binary-action model where Sender has maxmin preferences over Receiver types and maximizes the probability of inducing the favorable action. However, Receiver types represent an ambiguous posterior about a binary state of the world rather than a payoff-relevant characteristic which does not directly interact with beliefs about the state. This model captures substantively different applications—e.g., voters with common ideology who privately read outside news sources before listening to a politician's speech, rather than the equally-informed voters with different ideological positions in my model. Working with belief-independent Receiver types also means that I am able to characterize Receiver's posterior distribution and thus provide a sharp testable prediction—all posteriors in a known interior interval are equally likely. Methodologically, because my formulation features a simpler interaction between Receiver's type and Sender's signal, I am able to extend my approach to a continuous-state case.

A literature in robust mechanism design has also used moment conditions alongside maxmin preferences. Wolitzky [72] considers a bilateral trade model where each agent has

[1] Other works use maxmin preferences in Bayesian persuasion settings, but are much more distinct. In Kosterina [54], possible Receiver type distributions are distortions of a "reference distribution;" in Dworczak and Pavan [33], there is full ambiguity about Receiver's posterior belief; and in Laclau and Renou [55] and Beauchêne et al. [9], Receiver has maxmin preferences.

a valuation in $[0,1]$ and knows only the mean of the other agent's type distribution. In that model, agents' worst-case beliefs have binary support. Here, it is similarly possible to define a binary-support worst-case Receiver type distribution, but Sender's desire to induce indifference between many such distributions means the optimal posterior distribution has interval support. In Carrasco et al. [21], a principal with maxmin preferences offers a surplus-maximizing contract to a privately informed agent. Similar to my model, the agent's type distribution has known mean and support $[0,1]$. As in Hu and Weng [44] and my work, the optimal mechanism for the principal induces a payoff that is piecewise linear in the agent's type. Finally, Carrasco et al. [20] considers a setting where a seller with maxmin preferences faces an unknown distribution of buyer valuations. The seller knows the first $N-1$ moments of the valuation distribution and an upper bound on the Nth moment. Similar to the concavification argument of Kamenica and Gentzkow [48], optimal transfers are given by the non-negative monotonic hull of a degree-N polynomial.

1.3 Model

1.3.1 Setup and Preferences

There is one Sender (she) and one Receiver (he).[2] Both players share a common prior $F \in \Delta([0,1])$ about the state of the world $\omega \in [0,1]$, with $\mathbb{E}_F[\omega] = \pi \in (0,1)$. Only Receiver knows his private type $r \in [0,1]$, but the mean Receiver type $r^* \in (0,1)$ is common knowledge. In Section 1.6, I describe potential relaxations of these assumptions which endow Sender with less precise information about states or Receiver types.

Sender considers potential Receiver type distributions T in the set

$$\mathcal{T} = \left\{ \text{cdf T over } [0,1] \;\middle|\; \int r \, dT(r) = r^* \right\}.$$

I restrict Sender to the standard Bayesian persuasion tool of committing ex-ante to a Blackwell experiment, i.e., a state-dependent signal distribution, and in particular do not allow her to elicit Receiver's type in order to capture the public-communication interpretation of this model. After Sender communicates, Receiver chooses a binary action $a \in \{0,1\}$ whose utility depends on the state and on Receiver's type:

$$u_R(a, \omega, r) = a\,(\omega - r).$$

Thus when Receiver believes $\mathbb{E}[\omega \,|\, \text{Sender's message}] > r$, he strictly prefers $a = 1$, and when the opposite inequality holds he strictly prefers $a = 0$.[3] The explicit functional form used here is for ease of exposition only. Whenever Receiver's utility is a linear function of the state, his action depends only on the mean of his posterior belief about the state, and my results still hold (under an appropriate re-normalization of the interval of Receiver types).

[2] The presence of one Receiver with an unknown type may also be interpreted as a population of Receivers, each with a known type, with which Sender communicates publicly.

[3] Receiver's choice when indifferent will not affect equilibrium outcomes, but will affect Sender's equilibrium strategy. I discuss this tie-breaking issue in Section 1.4.1.

Sender's goal is to maximize the probability of inducing the high action $a = 1$ independent of the true state ω and true Receiver type r:

$$u_S(a, \omega, r) = a.$$

1.3.2 The Maxmin Persuasion Problem

Since Receiver's choice of action depends only on the mean q of the posterior belief distribution, I can follow Blackwell [13] and directly consider Sender choosing a distribution of posterior means G such that G is a mean-preserving contraction of the prior distribution F. The set of feasible distributions of posterior means is therefore

$$\mathcal{G} = \left\{ \text{cdf } G \text{ over } [0,1] \;\middle|\; \int_0^x G(q)\,dq \leq \int_0^x F(q)\,dq \;\; \forall \;\; x \in [0,1] \right.$$
$$\left. \text{and } \int_0^1 G(q)\,dq = \int_0^1 F(q)\,dq \right\}.$$

I follow the literature in referring to this constraint as Bayes-plausibility. Note that when $\text{supp}(F) = \{0, 1\}$, a case which I refer to as *binary support*, any posterior distribution that satisfies the equality at $x = 1$ satisfies the inequality for all $x \in [0, 1)$.

Using this formulation and Receiver's preferences, I rewrite Sender's utility as

$$u_S(q, r) = \mathbf{1}(q > r),$$

where I assume that an indifferent Receiver chooses Sender's less-preferred action, $a = 0$. Sender's full optimization problem is therefore

$$\max_{G \in \mathcal{G}} \left\{ \min_{T \in \mathcal{T}} \int \int \mathbf{1}(q > r)\,dG(q)\,dT(r) \right\}. \tag{1.1}$$

I state the optimization problem using a maximum and minimum, rather than supremum and infimum; the tie-breaking rule for indifferent Receivers ensures that the maximum and minimum are well-defined (see the proof of Lemma 1 in Appendix A.1.1 for details). The main difference from standard Bayesian persuasion with private information is the presence of an endogenously-determined Receiver type distribution.

1.3.3 MPC Games

I characterize the solution to the maxmin persuasion problem by reframing Sender's maxmin preferences as a zero-sum game, in which Sender designs a distribution of posterior means and then Nature adversarially designs a type distribution. More generally, my persuasion model can be viewed as a special case of a more general game which I call an *MPC game*. In this game, players $1, ..., N$ simultaneously[4] choose distributions $G_1, ..., G_N$ that are mean-preserving contractions of corresponding distributions $F_1, ..., F_N$. A realization x_i is drawn from each

[4]Simultaneous choice is a feature of most prior literature on games of this kind; I discuss in Section 1.4.1 why the switch from sequential to simultaneous moves does not affect the result.

distribution G_i to produce a vector of realizations $x = (x_1, ..., x_N)$. A prize allocation rule $A(x) : \mathbb{R}^N \to \mathbb{R}^N$ determines each player's payoff as a function of the realizations.[5] One simple prize allocation rule is to assign the player with the highest realization a payoff of 1 and all other players a payoff of 0; such a rule fits my model, where Sender gets a payoff of 1 if and only if the realized posterior exceeds the realized Receiver type. Various choices of F_i have been paired with this prize allocation rule in prior literature. In particular, as noted in the previous section, if a cdf F_i over a positive interval $[0, c]$ or over $\mathbb{R}_+$ has binary support, then any G_i with the same domain and mean as F_i is a mean-preserving contraction of F_i. MPC games where the F_i have binary support and domain $\mathbb{R}_+$ have been used to describe campaign spending or distribution of revenues by politicians [24, 64, 68]. Changing the domain to a finite interval $[0, c]$ has been used to model all-pay auctions with complete information [8, 43]. Further specifying the domain as $[0, 1]$ and potentially allowing F_i to have non-binary support can represent competition between different Senders attempting to persuade a single Receiver [6, 14, 45]. However, this work is the first to explicitly use the connection between MPC games and persuasion by a single Sender with maxmin preferences.

1.4 The Binary-State Setting

In this section, I fully characterize Sender's optimal distribution when the prior F has binary support, so that a distribution of posterior means is the same as a posterior distribution (I use the latter expression for simplicity). This case is equivalent to a 2-player MPC game where F_1 equals F, with domain $[0, 1]$, binary support, and mean π; F_2, which represents Nature's mean constraint, has domain $[0, \infty)$, binary support, and mean r^*.[6] The solution to the maxmin persuasion problem of Equation (1.1), as well as Sender's optimal posterior distribution under slight variations of my model, follows from extending earlier results about MPC games. Proposition 1 shows that when the prior π is weakly less than $1/2$, Sender uniquely selects an upper-truncated uniform distribution with an atom at $q = 0$. Under favorable tie-breaking—where an indifferent Receiver chooses $a = 1$ rather than $a = 0$—Corollary 1 in Appendix A.1.4 shows that Sender may modify this solution by also placing an atom at $q = 1$.

1.4.1 Maxmin Persuasion as an MPC Game

Equivalence of the maxmin persuasion problem in Equation (1.1) and the MPC game specified above rests on two results. The first (Lemma 1 in Appendix A.1.1) is that Nash equilibrium strategies for Sender in the MPC game are equivalent to optimal posterior distributions in the sequential-move game implied by Sender's maxmin preferences. This result follows from a minimax theorem in Fan [36], which shows that Sender's maxmin and minmax utilities are

[5] This setup may remind the reader of the literature on contests. In recent work, Antsygina and Teteryatnikova [2] also combines information design and contests, but focuses on a standard cost-of-effort setup for contest participants with a third party providing information about value profiles. In contrast, I represent an information design problem as a contest with Nature as a participant.

[6] This equivalence also holds when F does not have binary support. However, I do not use it in characterizing Sender's optimal distribution with a continuous-support prior in Section 1.5.

equal, and therefore equal to the utility from the MPC game.

The second result (Lemma 2 in Appendix A.1.1) is that tie-breaking against Sender is equivalent to ignoring tie-breaking but allowing the Receiver type distribution to be unbounded above.[7] Since Nature moves second, I use tie-breaking against Sender to ensure that the minimizing Receiver type distribution for each posterior distribution is well-defined; however, most results on MPC games use even tie-breaking, where a posterior $q = r$ convinces that Receiver type with probability 1/2. With unfavorable tie-breaking, in order to persuade Receiver type r, Sender must generate posterior $q_r^\varepsilon = r + \varepsilon$ for arbitrary $\varepsilon > 0$. Thus, unlike in an MPC game with even tie-breaking, Sender can never persuade type $r = 1$. I can restore the equivalence between maxmin persuasion and MPC games by allowing Nature in the MPC game to generate a Receiver type $r = 1 + \eta$ for arbitrary $\eta > 0$. Then Sender can attain her even tie-breaking utility for all interior Receiver types in the $\varepsilon \to 0$ limit and Nature can replace any instance of $r = 1$ with $r = 1 + \eta$ in the $\eta \to 0$ limit.

1.4.2 Optimal Posterior Distributions

Having established equivalence between the maxmin persuasion problem and an appropriate MPC game, the solution to the maxmin persuasion problem closely resembles Theorem 4 of Hart [43]. I extend that result by providing an alternative proof which shows uniqueness of Sender's optimal distribution when $\pi \leq 1/2$, as well as necessary and sufficient conditions on any maxmin-optimal posterior distribution when $\pi > 1/2$. Let δ_x be the Dirac distribution with all mass at $q = x$ and $U[x, y]$ be the uniform distribution over the interval $[x, y]$. The following describes Sender's optimal posterior distribution:

Proposition 1. *Let $supp(F) = \{0, 1\}$ and let ties be broken against Sender.*
If $\pi > 1/2$, then a posterior distribution G^* is optimal for Sender if and only if $\mathbb{E}_{G^*}[\omega] = \pi$ and $G^*(x) \leq x \ \forall \ x \in [0, 1]$.
If $\pi \leq 1/2$, then Sender's unique optimal posterior distribution G^* is as follows:

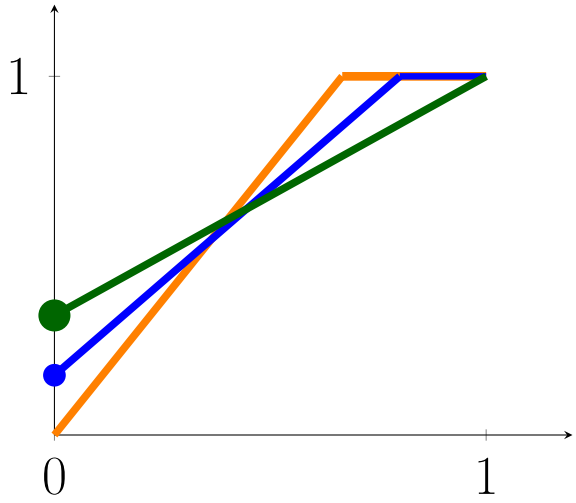
- *If $r^* \leq \pi \leq 1/2$,*
$$G^* = U[0, 2\pi].$$
- *If $\pi \leq r^* \leq 1/2$,*
$$G^* = \left(1 - \frac{\pi}{r^*}\right) \delta_0 + \frac{\pi}{r^*} U[0, 2r^*].$$
- *If $\pi \leq 1/2 \leq r^*$,*
$$G^* = (1 - 2\pi)\, \delta_0 + 2\pi\, U[0, 1].$$

Proof. See Appendix A.1.2. □

Because Nature can choose a binary support distribution where Receiver type $r = 0$ is always persuaded and Receiver type $r = 1$ is never persuaded, Sender's payoff cannot exceed

[7] Similar observations have been made in the context of all-pay auctions by Szech [70] and Gelder et al. [38]. The latter model finds that when ties may occur on intervals with positive measure, players' equilibrium strategies involve multiple disjoint intervals with an atom at 0, rather than the single interval and atom at 0 that arises in my model.

Figure 1.1: Optimal Posterior Distributions with Binary State



Several optimal posterior distributions in the binary-state setting for different values of the mean state π and mean Receiver type r^*.
In blue, $r^* = 1/4 < 1/3 = \pi$, and $G^* = U[0, 2/3]$.
In orange, $\pi = 1/3 < 2/5 = r^*$ and $G^* = (1/6)\,\delta_0 + (5/6)\,U[0, 4/5]$.
In green, $\pi = 1/3 < 3/5 = r^*$ and $G^* = (1/3)\,\delta_0 + (2/3)\,U[0, 1]$.

$1 - r^*$. Whenever Sender chooses a posterior distribution with a convex cdf, this Receiver type distribution is indeed optimal for Nature, and Sender attains her maximum payoff. Since the uniform distribution $U[0, 1]$ has the smallest mean among distributions with convex cdfs, this choice is feasible for Sender if and only if $\pi \geq 1/2$ (and multiple such distributions are feasible when $\pi > 1/2$). When $\pi < 1/2$, Sender chooses a distribution that is as close to uniform as possible given her Bayes-plausibility constraint. This choice requires her to place an atom at $q = 0$, truncate the upper bound of the distribution's support below $q = 1$, or both. Fixing $\pi \leq 1/2$, for small r^* Sender truncates the support at 2π but places no atom at 0. As r^* increases, Sender simultaneously increases the size of the atom and moves the upper bound of the support towards 1. A higher average Receiver type makes high posteriors more valuable to Sender, so she is willing to sometimes fully reveal the low state in order to generate more of these posteriors. Figure 1.1 shows three examples of optimal posterior distributions, corresponding to the three cases of Proposition 1.

Hart [43] does not establish uniqueness of this Nash equilibrium of the MPC game. A related work, Amir [1], shows through explicit calculations of players' utilities under various distributions that Sender's Nash equilibrium strategy is unique when $\pi \leq 1/2$ (Theorems 4 and 5 in that work) but gives only a partial characterization of optimal strategies for Sender when $\pi > 1/2$ (Theorem 10 in that work). In the maxmin persuasion setting with tie-breaking against Sender, I am able to avoid issues with limits of ε-approximating distributions and close that gap: Lemma 4 in Appendix A.1.2 gives a necessary and sufficient condition for Sender's optimal distribution when $\pi > 1/2$. Additionally, in Lemma 8 of Appendix A.1.3, I provide a novel geometric proof of Sender's optimal posterior distribution for the case $\pi \leq 1/2$, including its uniqueness, which leverages the concavification approach of Kamenica

and Gentzkow [48]. This proof informs my approach in the continuous-state setting.

1.5 The Continous-State Setting

While the solution when F has binary support is especially sharp, that restriction may not always be plausible. In this section, I consider the maxmin persuasion problem of Equation (1.1) when F is a continuously differentiable and unimodal cdf over $[0,1]$ with $F(0)=0$. I assume that, for some mode $m \in [0,1]$, the density f is strictly increasing on $[0,m)$ and strictly decreasing on $(m,1]$. To rule out the binary-state solution, I also assume that $f'(0) < 1-2\pi$. In this setting, a double-truncated uniform distribution of posterior means is optimal when r^* is sufficiently small (Proposition 2) or large (Proposition 3). Analogous results hold when, rather than being continuous, F is supported on N values $\{q_1, ..., q_N\} \in [0,1]^N$ with $N > 2$; I provide full details and proofs of this extension in Appendix A.2.6. Before turning to the main result, I first discuss two simple cases which extend the intuitions of the binary-state setting.

1.5.1 Simple Continuous Priors

In the continuous-state setting, Sender's constraint is different from Nature's. It is no longer true that any cdf G with support $[0,1]$ and mean π is a mean-preserving contraction of the prior F. Instead, the chosen cdf must additionally satisfy the *integral constraint*

$$\int_0^x G(q)\,dq \leq \int_0^x F(q)\,dq \;\;\forall\;\; x \in [0,1].$$

Under the assumptions $F(0)=0$ and $f'(0) < 1-2\pi$, this constraint prevents Sender from choosing any of the optimal distributions in Proposition 1, as they violate it in the interval $(0,\varepsilon)$ for $\varepsilon > 0$ sufficiently small.

Despite this new constraint, two cases of the continuous-state model are easy to solve using the intuitions of the previous section. Nature may still choose the binary support distribution which generates only Receiver types $r=0$ and $r=1$, so Sender's utility is still upper-bounded by $1-r^*$. Thus for any F that first-order stochastically dominates $U[0,1]$—so that $F(q) \leq q \;\;\forall\;\; q \in [0,1]$—it is easy to see that full disclosure is optimal, since it ensures that $G^* = F$ and Sender's utility attains the upper bound. This condition generalizes the case where F is unimodal with $m=1$.

When F is concave but not uniform, it must be that F lies strictly above $U[0,1]$ on $(0,1)$ and therefore that $\pi < 1/2$. Additionally, the uniform distribution $U[0,2\pi]$, which was uniquely optimal when $r^* \leq \pi \leq 1/2$ in the binary-state setting of Proposition 1, satisfies the integral constraint. To see why, note that the shape of F ensures that $U[0,2\pi]$ lies strictly below F on $(0,c)$ for some $c < 1$; therefore the integral constraint is satisfied with equality at $x=0$ and strict inequality for $x \in (0,c]$. The difference between the left- and right-hand sides of the constraint strictly decreases for $x \in (c,1)$, but only reaches 0 at $x=1$: thus the weak inequality is preserved on the entire interval $[0,1]$.[8] The binary-state maxmin persuasion

[8]This geometric approach to the integral constraint will be key in proving Proposition 2.

problem is a relaxed version of the continuous-state maxmin persuasion problem, so a feasible optimal solution for Sender in the former must be optimal in the latter. Thus, if $r^* \leq \pi$ and F is concave but not uniform, then $G^* = U[0, 2\pi]$ is the unique optimal distribution for Sender. This condition generalizes the case of unimodal F with $m = 0$.

In the maxmin setting, Nature's mean constraint represents information Sender possesses which allows her to consider only a particular set of possible Receiver type distributions. Thus it is reasonable for Nature to face only a mean constraint while Sender also faces the integral constraint. However, in the case of competitive persuasion, where both parties are persuading Receiver about a common state, it is natural to require all players to choose mean-preserving contractions of the same continuous prior. This case is studied in Hwang et al. [45], where the optimal distribution of posterior means divides the prior support into finitely many intervals and alternates between matching the prior and generating a linear mean-preserving contraction on each interval. Nature's weaker constraint in my setting rules out this result.

1.5.2 Optimal Distributions with a Small Mean Receiver Type

Towards providing sufficient conditions for a double-truncated uniform distribution (henceforth DTU) to be optimal for Sender, I first establish notation. A DTU places no mass on any $q \in [0, \ell]$, an atom at $q = \ell$, uniform mass on all $q \in [\ell, u]$, and no mass on any $q \in [u, 1]$. It thus has three parameters: the lower truncation length, the size of the atom at ℓ, and the upper truncation length. Because any Bayes-plausible DTU's mean must be π, the atom size is uniquely determined by the truncation lengths. That is, given $[\ell, u] \subseteq [0, 1]$ there is only one DTU with mean π and support $[\ell, u]$. I can thus characterize a DTU by the slope $\beta > 0$ and intercept $y \in [0, 1)$ of the uniform portion of its cdf, writing it as G_y^β. Explicit formulas for the relationship between truncation lengths, atom size, and slope, as well as bounds on these parameters, are in Appendix A.2.1. For each y, there is a slope $\beta(y)$ which delivers Sender's highest utility among DTUs with intercept y (Lemma 9 in Appendix A.2.2); I refer to the DTU $G_y^{\beta(y)}$ as y-optimal. Let $q_i(\beta, 0)$ be the smallest nonzero point of intersection between the DTU G_y^β and the prior F. Figure 1.2 shows an example of the 0-optimal DTU when F is a truncated normal distribution, highlighting the notation above. With notation fixed, the following proposition describes Sender's choice for small r^*:

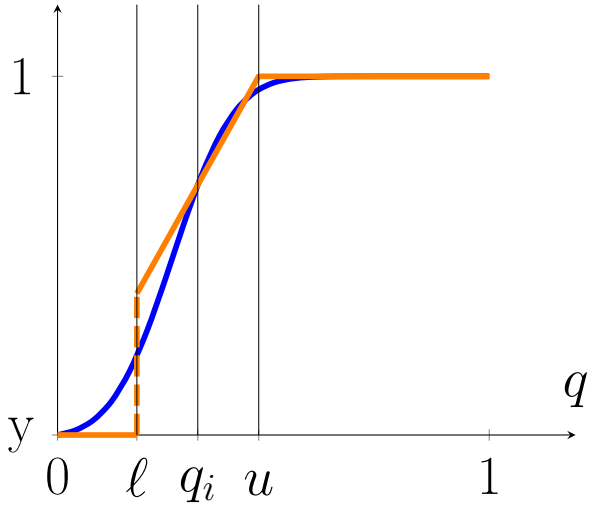
Proposition 2. *Let $r^* \in (0, q_i(\beta(0), 0)]$. The 0-optimal double-truncated uniform distribution, $G_0^{\beta(0)}$, attains Sender's highest possible utility. Any other distribution which attains that utility must have the same concavification as $G_0^{\beta(0)}$.*

Proof. See Appendix A.2.5. □

The key step in the proof is to show that the integral constraint binds only at a single interior point, the intersection $q_i(\beta(0), 0)$ between the DTU and the prior cdf (Lemma 11 in Appendix A.2.3).[9] Using this result, I can adapt the the strategy used in my geometric proof

[9]When F has finite support, the integral constraint need not bind at any interior points. In Appendix A.2.6, I show that if it does bind, it must do so only on a subset of $\text{supp}(F)$. Letting $q_{\min}$ be the minimal element of $\text{supp}(F)$ where the integral constraint binds, Proposition 2 holds after replacing $q_i(\beta(0), 0)$ with $q_{\min}$.

Figure 1.2: Key Values of a Double-Truncated Uniform Distribution



An example 0-optimal DTU, $G_0^{2.15}$, in orange. The prior F (blue) is a $N(1/5, 1/3)$ distribution truncated in $[0, 1]$. Its support is $[\ell, u] = [0.18, 0.47]$; its intercept is $y = 0$; its slope is $\beta(0) = 2.15$; the smallest nonzero point of intersection between the DTU and the prior is $q_i(\beta(0), 0) = 0.32$.

of Proposition 1. Towards simplifying the integral constraint, note that a DTU G_0^β will have zero, one, or two interior intersections with F depending on its slope.[10] Writing the integral constraint as a function of x,

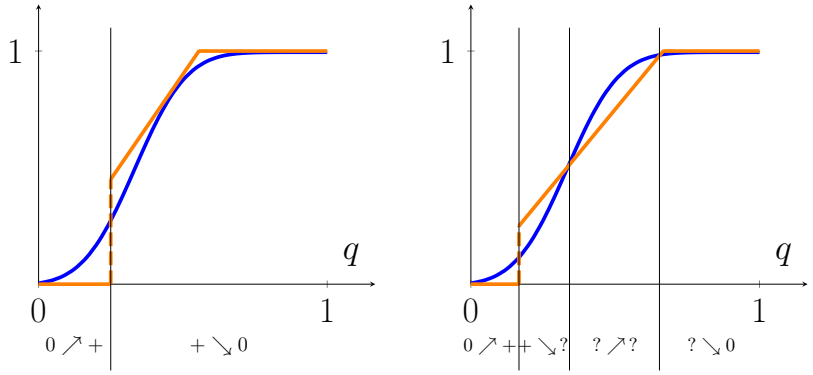
$$v(x) = \int_0^x F(q)\,dq - \int_0^x G_0^\beta(q)\,dq,$$

the intersections of F and G_0^β can be used to infer whether v is increasing or decreasing on particular intervals. Figure 1.3 shows an example of this approach. Combined with the observation that $v(0) = v(1) = 0$, this behavior allows me to show that if G_0^β has two interior intersections with F, then it satisfies the integral constraint if and only if $v(q_i(\beta, 0)) \geq 0$. To select among these Bayes-plausible DTUs, note that when $y = 0$ the slope of a concavified DTU G_0^β equals β in the lower truncation interval $[0, \ell]$ and uniform interval $[\ell, u]$. Thus Sender's 0-optimal DTU is given by making β small (to minimize Nature's utility) while satisfying the simplified integral constraint.

The remainder of the proof uses the fact that the integral constraint binds only at $q \in \{0, q_i(\beta(0), 0), 1\}$ for $G_0^{\beta(0)}$ and adapts the strategy used in my geometric proof of Proposition 1. First, I show that any DTU which delivers higher utility for Sender than $G_0^{\beta(0)}$ cannot be Bayes-plausible. I then approximate an arbitrary optimal distribution H by a DTU, using the concavification of H to ensure that this DTU upper-bounds H everywhere above the lower truncation. If H delivers Sender strictly higher utility than $G_0^{\beta(0)}$, the approximating

[10] Slight variations of these cases may occur and are dealt with in the proofs of Appendix B, but can be ignored to simplify the intuition.

Figure 1.3: The Sign of the Integral Constraint



Two DTUs with intercept 0 and varying slopes (orange) and their relationship to the prior F (blue). The guidelines show where the function v, which represents the difference between the integral of F and that of the corresponding DTU, switches from increasing to decreasing or vice-versa.

DTU must do so as well; therefore it cannot be Bayes-plausible, and neither is H itself. If H delivers Sender the same utility as $G_0^{\beta(0)}$, then the approximating DTU is precisely $G_0^{\beta(0)}$ and the concavification of H equals that of $G_0^{\beta(0)}$. This last step relies crucially on the slope of $G_0^{\beta(0)}$ being the same as that of its concavification in the lower truncation interval $[0, \ell]$. For a DTU with intercept $y > 0$, this property will no longer hold, and as a result the concavified optimal distribution will no longer be unique.

While the concavification of $G_0^{\beta(0)}$ is the unique concavified distribution that maximizes Sender's utility, $G_0^{\beta(0)}$ is not itself a unique solution to the maxmin persuasion problem. In the binary-state setting, Sender's optimal distribution was equal to its concavification everywhere on $[0, 1]$, and any other distribution with the same concavification would have a different mean. In the continuous-state setting, a DTU differs from its concavification on the lower truncation interval, so it is possible for a non-DTU distribution H to have the same mean and concavification as a DTU. Thus uniqueness of the concavification is the strongest result that can be obtained.

1.5.3 Optimal Distributions with a Large Mean Receiver Type

When $r^* > q_i(\beta(0), 0)$, characterizing both the optimal DTU and optimal distributions more generally becomes more difficult. In fact, an optimal distribution may not exist, though Sender's supremum utility over a sequence of distributions converging to optimality is always well-defined. Despite these challenges, I can still show that for sufficiently large values of r^*, DTUs are not dominated by other distributions. This result holds without alteration when F has finite support:

Proposition 3. *Let $r^* \in [\pi, 1)$. Then no distribution of posterior means gives Sender strictly higher utility than all DTUs.*

Proof. See Appendix A.2.5, or Appendix A.2.6 for the finite-state case. □

This proof is similar in approach to that of Proposition 2. When $y > 0$, the slope of a concavified DTU G_y^β is larger in the lower truncation interval $[0, \ell]$ than that of the DTU itself (since the concavification passes through the origin, while the DTU has intercept y), but is again equal to β in the uniform interval $[\ell, u]$. Thus the concavified DTU has a kink at $q = \ell$. However, setting $r^* \geq \pi$ ensures that the kink does not affect the value of the concavified DTU at r^*. Then, as in Proposition 2, Sender's y-optimal DTU for each intercept y is given by minimizing the slope β subject to the integral constraint. Unlike in that proposition, I cannot directly characterize which choice of intercept is optimal. In fact, since the set $[0, 1)$ of possible intercept choices is not compact, it may be that the optimal choice is $y = 1$ and Sender's highest utility is attained only in the limit. However, I can still use the fact that each y-optimal DTU has minimal slope among Bayes-plausible DTUs with intercept y to extend the bounding argument of Proposition 2. This approach rules out as infeasible any distribution that delivers strictly higher utility than all DTUs, but again leaves room for non-DTU distributions that attain Sender's highest possible utility.

1.5.4 Non-Uniform Optimal Distributions

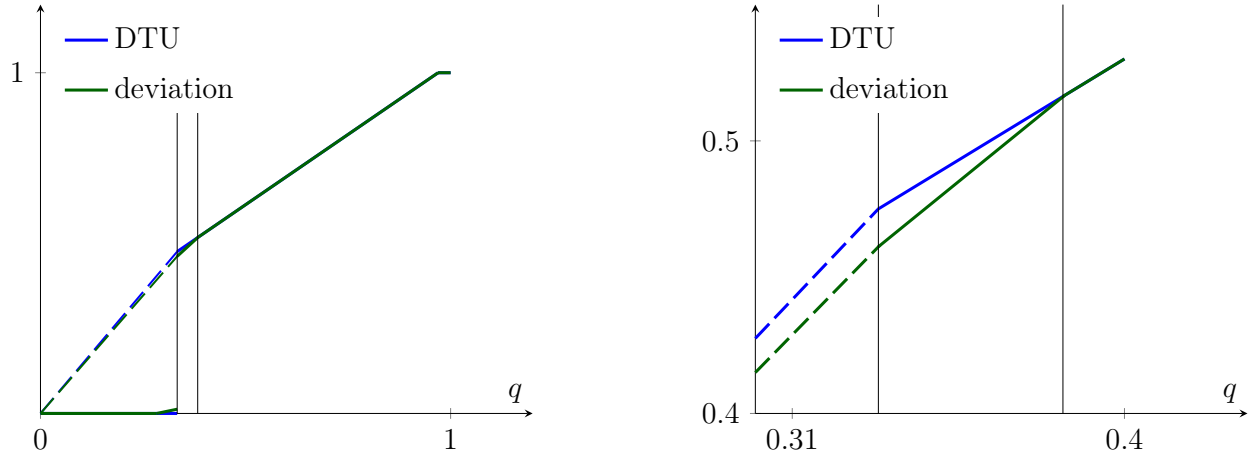
The result of Proposition 2 provides an appealing reason for focusing on DTUs as opposed to other maxmin-optimal posterior distributions: outside of the lower-truncation region, the optimal DTU is precisely equal to the unique optimal concavification. However, in the large-r^* case of Proposition 3, the optimal concavification is no longer unique. To see why, assume an optimal DTU with $y > 0$ exists. Its concavification passes through the origin rather than the point $(0, y)$, so it has a kink at $q = \ell$. This kink can be used to alter the DTU without affecting Sender's utility. In particular, consider a distribution that places slightly positive mass in the interval $[\ell - \varepsilon, \ell)$, has a smaller atom than the DTU at $q = \ell$, and places slightly less mass than the DTU in the interval $(\ell, \ell + \varepsilon]$. This distribution, shown in Figure 1.4, changes slope at ℓ and $\ell + \varepsilon$, but is equal to the DTU for $q \notin (\ell - \varepsilon, \ell + \varepsilon)$. Whenever $r^* \geq \ell + \varepsilon$, and in particular when $r^* \geq \pi$ (the case in Proposition 3) the deviation delivers the same utility for Sender.

1.5.5 Studying an Intermediate Mean Receiver Type

The deviation in Figure 1.4 also sheds light on the difficulty of characterizing the optimal distribution when $r^* \in (q_1(\beta(0), 0), \pi)$. In the binary-state setting the optimal distribution equals its concavification. With a continuous state, lower truncation is one possible response to the integral constraint, but it is not a unique solution for Sender because that constraint may only bind at a finite set of interior points. For instance, the deviation in Figure 1.4 gives Sender a greater utility than the corresponding DTU when $r^* \in (0, \ell + \varepsilon)$ and is feasible whenever the integral constraint does not bind in that interval. Without further structure on the space of possible deviations from DTUs, even numerical approaches with a parametric prior distribution provide no insight, since they would require a novel algorithm to search over all mean-preserving contractions of the prior.

Despite this challenge, I am able to shed light on the prevalence of the intermediate-r^* case by numerically estimating $q_i(\beta(0), 0)$ within a class of parametric prior distributions. For truncated normal priors—generated by taking a $N(\mu, \sigma^2)$ distribution and truncating it

Figure 1.4: Deviations from Double-Truncated Uniform Distributions



A potential deviation (solid green) from a DTU (solid blue); the concavification of each distribution is shown by dashed lines of the same color. The second panel focuses on the concavified deviation's double kink (at both guidelines), while the concavified DTU's has only one kink (at the first guideline).

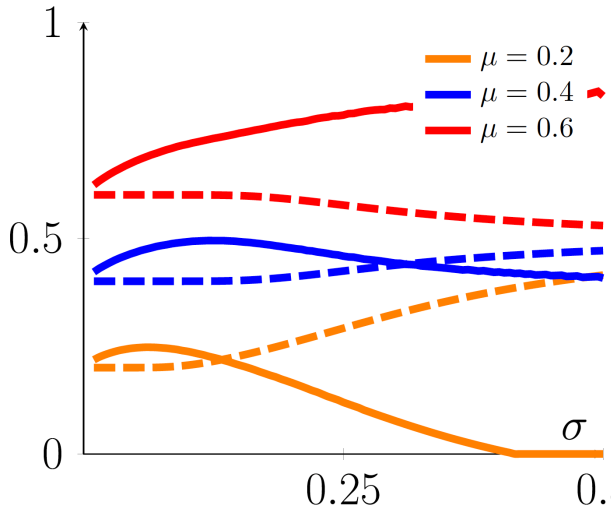
to lie in the unit interval—I show numerically that there is a gap between Propositions 2 and 3 only when $\mu < 1/2$ and σ is large enough. For example, when $\mu = 0.2$, shown in orange in Figure 1.5, there is a gap only when $\sigma \geq 0.135$.

Full details of the algorithm for computing $q_i(\beta(0), 0)$ are in Appendix A.3; Figure 1.5 shows an example of the output from these computations. Each color represents a fixed mean μ of the generating normal distribution, with the x-axis representing that distribution's standard deviation σ. Because the normal distribution is truncated to produce a prior in $[0, 1]$, the "true mean" π of that prior depends on both μ and σ; it is shown as a dashed line. The solid line shows the numerically computed value $q_i(\beta(0), 0)$. Thus there is a gap between the small-r^* case of Proposition 2 and the large-r^* case of Proposition 3 if and only if a dashed line lies above its corresponding same-color solid line. When $\mu > 1/2$, this property never holds and there is no gap between Proposition 2 and Proposition 3. When $\mu < 1/2$, there is no gap for σ small enough, but a gap arises for larger σ. However, making σ too large violates the assumption $f'(0) < 1 - 2\pi$, invalidating the propositions. These results suggest that double-truncated uniform distributions are optimal for many possible priors.

1.6 Extensions

In the motivating example, a politician has a well-defined prior belief about the state of the world and knows the average voter's cost-effectiveness threshold, but she makes no further assumptions on the distribution. This difference in information is not unreasonable: the politician can fine-tune the details of her welfare program, but voter preferences are subject to a number of factors outside her control, e.g., opposition campaigning and news coverage. Loosely speaking, limited data about voter preferences allows the politician to estimate the

Figure 1.5: Numerical Results for Intermediate Receiver Types



Numerically computed values of $q_i(\beta(0), 0)$ and their corresponding prior means π using a $N(\mu, \sigma^2)$ distribution truncated to have support $[0, 1]$. For each color, the solid line is $q_i(\beta(0), 0)$ and the dashed line is the prior mean π after truncation, as functions of the generating variance σ.

population mean with convergence rate $1/n$, but estimation of the distribution (or any given quantile) converges at rate $1/\sqrt{n}$; she may thus be more willing to base her strategy on the former than the latter. Despite these justifications, it may still be realistic to weaken these informational assumptions; I do so in this section and discuss how my existing results extend.[11]

1.6.1 Doubly-Maxmin Sender Preferences

I first consider a Sender who knows only the mean state and mean Receiver type, with maxmin preferences over all possible pairs of independent distributions fitting those moment restrictions. In the binary-state setting, the mean state and state distribution are equivalent. In the continuous-state case, allowing Nature to choose both distributions in response to Sender's choice of information structure makes Sender's optimization problem ill-posed since her set of feasible distributions depends on the prior distribution. There are two straightforward ways to ensure this set is well-defined. First, I can alter the order of the moves: Nature chooses a unimodal prior distribution (as described in Section 1.5), then Sender chooses a distribution of posterior means, and finally Nature chooses a Receiver type distribution with the known mean. In this case, my characterizations of Sender's optimal distribution (Propositions 2 and 3) apply to each possible prior. The overall solution follows by minimizing (over possible priors) Sender's maxmin utility from the existing model. Given the lack of closed-form solutions in Section 1.5, this minimization requires a numerical approach. Second, I can ignore the prior altogether by letting Sender choose a distribution of posterior

[11] I am grateful to an anonymous referee for suggesting these approaches.

means, then let Nature simultaneously choose any mean-preserving spread of that distribution (which I can call a "prior") and any Receiver type distribution with the known mean. Since only the distribution of posterior means affects Sender's utility, this approach reduces to the binary-state setting of my model where Sender faces a mean constraint.

1.6.2 Flexible Mean Receiver Types

If Nature may freely select from among different mean Receiver types, the highest permissible mean Receiver type is always worst for Sender. However, I can partially relax the mean restriction by allowing Nature to choose higher mean Receiver types only by paying some cost (similar to the variational preference setup in Maccheroni et al. 61). If Nature is restricted to choosing a single mean Receiver type, then the results of my model are unchanged for that fixed mean and its value for Nature can be computed either analytically using Proposition 1 for the binary-state case or numerically in the continuous-state case. If Nature may randomize over mean Receiver types, then I can solve the binary-state case using the characterization in Proposition 1.

Let $u_s(r^*, r)$ be Sender's utility if she chooses the optimal distribution for mean Receiver type r^*, but in fact faces a realized mean Receiver type r (drawn from some distribution chosen by Nature). For any r^*, the value of $u_s(r^*, r)$ is given by the concavification of a truncated uniform distribution (following Proposition 1 and Lemma 3), so it is linear in r.[12] Thus if the expected mean Receiver type chosen by Nature is r^{**}, Sender's utility from choosing the r^*-optimal distribution is $u_s(r^*, r^{**})$ regardless of the full distribution of mean Receiver types. This expression is maximized by choosing the r^{**}-optimal posterior distribution, in which case Sender could do no better even if she knew the mean Receiver type was r^{**} with certainty. Therefore, fixing some zero-cost mean Receiver type r^*, Nature's gain from any distribution with mean Receiver type r^{**} is captured by the difference between Sender's maxmin utility at r^* and her maxmin utility at r^{**}. Sender's maxmin utility is a well-defined, continuously differentiable function of the mean Receiver type (computed using Proposition 1), so Nature's choice of r^{**} can be straightforwardly found by, e.g., setting the marginal cost of increasing the expected mean Receiver type equal to the marginal decrease in Sender's utility. Sender's optimal distribution when facing this new mean Receiver type is given by using r^{**} in Proposition 1.

In the continuous-state case, the presence of multiple optimal distributions (some of which do not have linear concavifications), the kink in the concavification of a double-truncted uniform distribution (as discussed in Section 1.5.4), and the potential gap between Propositions 2 and 3 rule out the approach above. Finding Sender's optimal response in this case would require a different characterization of her maxmin utility, so I leave it for future work.

[12] It is linear if r is in the support of Sender's optimal distribution. Any mass strictly above the support brings Nature no additional benefit, so I assume without loss that this choice is never made.

1.7 Conclusion

Bayesian persuasion provides a tractable model of communication that can be extended to include rich uncertainty about the Receiver who is the target of persuasion. This work contributes to a growing literature that also introduces ambiguity by posing a maxmin persuasion problem, where the Sender seeks to be robust to any possible prior belief about Receiver types with a known mean. In a binary-state setting, I show a connection to mean-preserving contraction (MPC) games, where competing players choose mean-preserving contractions of probability distributions to obtain the highest realization, and fully characterize Sender's optimal distribution. As in many other MPC games, when her constraint is strong enough Sender chooses a uniform distribution mixed with atoms at the lowest and highest posterior beliefs. These results highlight the importance of the tie-breaking assumption in persuasion problems and emphasize the strength of the maxmin criterion, which delivers strictly lower utility for Sender than any prior belief over Receiver types when the probability of the high state is less than $1/2$. I then use a geometric approach to show, in both a finite-support setting and a novel continuous-state setting, that uniform distributions with an atom at the lower bound of their support are in many cases still optimal. Unlike in the binary-state setting, these distributions now have support in the interior of $[0, 1]$, and Sender's optimal distribution is no longer unique. However, the intuition of linearizing the prior belief over states in order to make Nature indifferent between many worst-case Receiver type distributions is preserved.

Chapter 2

Discovery through Trial Balloons

A principal and an agent face symmetric uncertainty about the value of two correlated projects for the agent. The principal chooses which project values to publicly discover and makes a proposal to the agent, who accepts if and only if the expected sum of values is positive. We characterize optimal discovery for various principal preferences: maximizing the probability of the grand bundle, of having at least one project approved, and of a weighted combination of projects. Our results highlight the usefulness of trial balloons: projects which are ex-ante disfavored but have higher variance than a more favored alternative. Discovering disfavored projects may be optimal even when their variance is lower than that of the alternative, so long as their disfavorability is neither too large nor too small. These conclusions rationalize the inclusion of controversial policies in omnibus bills and the presence of moonshot projects in organizations.

JEL Classification: D21, D83
Keywords: trial balloons, public discovery, constrained information design

Chapter Acknowledgments: We thank Robert Gibbons, Stephen Morris, Alexander Wolitzky, and participants in the MIT Theory Lunch for helpful discussions and comments. Thanks especially to Roi Orzach for helping with an earlier version of this project. We acknowledge support from the National Science Foundation Graduate Research Fellowship under Grant No. 1745302.

2.1 Introduction

Many large proposals are composed of multiple smaller projects. For example, the 2022 omnibus appropriations bill passed in the US Congress includes funding for various federal agencies, transportation-related earmarks, and emergency aid for Ukraine;[1] a grant proposal may contain funding for multiple research designs carried out by a particular lab; or a worker may request their manager's approval for a slate of several projects rather than presenting them one-by-one. In these examples, and many others, even if the value of each *project*, or component issue in the overall *proposal*, is uncertain, there may be a known correlation between projects. After all, an appropriations bill reflects the priorities of the governing party; a scientific lab usually has a well-defined area of expertise; and a worker has a particular role or set of responsibilities. Given the complexity of the full proposal, it both unsurprising and empirically common for proposers in these settings to concentrate attention on one "headline" project before finalizing the proposal. For instance, the Biden administration campaigned in large part on infrastructure investment, then included infrastructure-related items like clean energy spending alongside changes to Medicare and the tax code in the Inflation Reduction Act.[2] The choice of which project to focus on thus becomes an important tool for designing such proposals.

We model this setting by considering a principal and an agent who face uncertainty about the agent's value for two correlated projects. The principal chooses which project value, if any, to publicly discover, and then proposes a set of projects to the agent. The proposal is enacted if and only if the agent believes the expected sum of project values is positive. We characterize optimal discovery for several natural forms of the principal's preferences—maximizing the probability of the grand bundle, of having at least one project approved, and of a weighted combination of projects. Our results highlight the usefulness of *trial balloons*: projects which are ex-ante disfavored but have higher variance than the alternative. In contrast to earlier literature, which generally finds benefits from *building consensus* through information about ex-ante favored projects, in our setting trial balloons have important benefits driven by potential ex-post disagreement on the approver's favored project. Even a disfavored project with lower variance than the more-favored alternative may be preferred if it is moderately disfavored or the prior is highly precise.

The trade-off between trial balloons and building consensus is driven by two forces: *spoiling* and *improvement*. Spoiling is the chance that, by obtaining an unfavorable draw, the principal will lose out on the sure-thing payoff that could have been obtained without generating any information. Improvement is the converse—a favorable draw may raise the agent's belief sufficiently to get ex-ante disfavored projects approved. Trial balloons have a weaker spoiling effect than other discovery rules, though it may be that spoiling is still likely enough for the principal to prefer no discovery at all. Full discovery—if permitted—generally has the strongest improvement effect because it introduces the most variance and thus the highest chance of a sufficiently positive realization. Our main results compare the relative magnitudes of these effects to characterize when trial balloons are optimal. We also provide a partial characterization of optimal discovery when the disfavored project has lower variance,

[1] As described by the National Conference of State Legislatures here.

[2] The bill's contents are summarized in a White House press release here.

as well as a partial comparison of single-project discovery to full discovery.

Using these characterizations, we describe rich comparative statics of the principal's optimal discovery rule. Because higher variance results in a stronger improvement effect, discovering a trial balloon is always better than discovering a consensus project. However, the effect of a more precise prior on optimal discovery depends on how much the principal values having the trial balloon approved. A trial balloon which is less important than the consensus project benefits from a more precise prior, in the sense that the principal prefers it to no discovery even at higher magnitudes of disfavorability. The opposite is true for a more important trial balloon: the range of magnitudes where the principal discovers it shrinks. Both results are driven by the rates at which spoiling and improvement effects vanish as the prior's precision increases. When the disfavored project has lower variance—and is therefore no longer a trial balloon—these results are altered by the consensus project's larger improvement and spoiling effects. In the limit, the disfavored project is discovered only when the magnitude of its disfavor is intermediate. If it is too disfavored, the principal prefers no discovery; if it is only mildly disfavored, the principal prefers discovering the consensus project. Optimal discovery away from the high-precision limit is qualitatively opposite to the trial balloon case. When the principal places more weight on the low-variance disfavored project, the intermediate favorability range where it is discovered grows as prior precision increases; less weight causes it to shrink with higher precision. These sharp differences based on the observable features of the projects available to the principal provide clear testable predictions.

Finally, we consider extensions of our stylized model to better reflect real-world proposals. Although we are motivated by the idea of a constrained principal who cannot fully design information about the projects, the restriction to discovery of a single project is stark: we relax this choice by allowing either one-shot discovery of both projects or a sequential discovery choice. We also consider noisy discovery, where rather than generating conclusive evidence about the project's value to the agent, the principal may only be able to generate a noisy signal. Lastly, we discuss which results remain tractable in a setting with N projects rather than only two. Our qualitative results about the settings in which trial balloons dominate due to the presence of spoiling effects extend, and emphasize how the desire to keep a consensus option on the table may lead to a focus on disfavored projects.

2.2 Related Literature

Our work relates to a broad literature on soliciting and using advice from potentially biased experts. Caillaud and Tirole [18] and Dziuda [34] describe hard evidence setting where a principal's disclosure rule is designed to "build consensus" by suggesting to the agent that some underlying state of the world is favorable. Our model focuses instead on the interplay between different dimensions of a state. Perhaps closest to our work is Brocas and Carillo [17], where a leader can decide whether or not to publicly gather information; this setup resembles a sequential discovery version of our model, but where the "dimensions" are signals of the same payoff-relevant state. The richer interaction between dimensions in our model means that, unlike in that work, one-shot discovery is a nontrivial problem in our setting.

A literature in political economy describes "policy bundling," or combining different

projects into a single proposal that receives an up-or-down vote. Many of these models are dynamic and study how today's project choice influences tomorrow's bargaining procedure or options.[3] Our model is instead a one-shot game, though we briefly discuss sequential discovery in Section 2.5.2. Another key difference is that rather than using a consensus policy featuring full preference alignment and an ideological dimension featuring disagreement, as is the case in Jackson and Moselle [46] and Chen and Eraslan [27], we allow for flexible levels of alignment. Indeed, our main results (Theorems 1 and 2) highlight how the ratio of agreement to disagreement shapes the principal's optimal discovery rule. Closer to our work is Câmara and Eguia [19]; however, that model focuses on how to bundle policies (using an information design setup) rather than how to communicate about an exogenous policy bundle.

Although we frame our model as one of symmetric uncertainty, where the principal must publicly *discover* information rather than committing to a *disclosure* rule for private information, this distinction is one of interpretation only. The principal's problem can also be viewed as a constrained version of the standard Bayesian persuasion problem (Kamenica and Gentzkow 48, with extensions surveyed by Kamenica 47) in which the principal chooses a signal structure to maximize a function of the agent's posterior beliefs. Our first constraint only allows the principal to disclose a single dimension of the multidimensional state, capturing feasibility constraints on the principal (such as the ability to generate only one prototype) as well as attention constraints on the agent (such as the ability to focus on only one spending plan in a large budget bill). Of note is a recent result by Malamud and Schrimpf [63], which shows that with a high-dimensional state the optimal unconstrained disclosure rule involves dimension reduction. We view this result as complementary, emphasizing the value of focusing on dimensionally-constrained information design. Our second restriction is that, for the chosen dimension, the principal must choose from an exogenously given set of signal structures rather than having access to any Blackwell experiment. Prior work on constrained information design has considered differently-structured limitations;[4] the closest to our restrictions is the general framework in Babichenko et al. [7]. However, that work focuses primarily on high-level characterizations and computational results rather than a tight description of optimal discovery rules.

2.3 Model

There are two projects $i \in \{1, 2\}$, which are relevant to both a principal (she) and an agent (he). Let $v = (v_1, v_2) \in \mathbb{R}^2$ represent the agent's value for each project, and $w = (w_1, w_2) \in [0, 1]^2$ with $w_1 + w_2 = 1$ be the relative weights the principal places on them. The principal and the agent share a common prior F from which v is drawn,

$$F \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right),$$

[3] For example, Chen and Eraslan [28], Fershtman [37], Lee [58], and Bils and Izzo [12].

[4] Le Treust and Tomala [57] studies a principal facing many copies of the same persuasion problem, whereas Tsakas and Tsakas [71] considers a single problem with noisy communication.

so that the marginals F_1 and F_2 are normally distributed with variances σ_1^2 and σ_2^2 respectively, and have correlation $\rho \in (0, 1)$.[5] We now describe the timing of the discovery game, noting variations as they arise:

1. The principal chooses whether to discover v_1, v_2, or neither.[6]
2. The value v_i of the chosen project, if any, is publicly revealed.[7] The agent forms a posterior belief G about v by using Bayes's Rule.
3. The principal makes a proposal by selecting a subset of projects S. This subset need not contain the discovered project (and also need not exclude it).
4. The agent approves the proposal so long as $\sum_{i \in S} \mathbb{E}_G[v_i] \geq 0$.
5. The principal receives $\sum_{i \in S} w_i$ if her proposal is approved, and 0 otherwise.[8]

This formal description allows us to make precise the language of trial balloons and consensus projects from the introduction. Using a *trial balloon* means discovering a project i with $\mu_i < \mu_j$ because project i is ex-ante less favored by the agent, as well as $\sigma_i \geq \sigma_j$ because the trial balloon's value is more uncertain. We refer to a project i with $\mu_i < \mu_j$ without restriction on its variance as *disfavored* and call project j a *consensus project*. This description is a slight misuse of notation because it may be that $\mu_j < 0$ and both projects are ex-ante disfavored, but we will focus especially on the case where $\mu_i < 0 < \mu_j$, so that the consensus project is, as the name suggests, ex-ante open to approval by the agent. This case allows us to clearly contrast the spoiling and improvement effects discussed in the introduction.

2.4 Single-Project Discovery

In this section, we fully characterize the cases in which discovering the value of a trial balloon is best for the principal. We also describe optimal discovery of disfavored projects even when their variance is lower than that of the favored alternative. The principal's decision in the single-project discovery setting is fully captured by the spoiling versus improvement trade-off. Both effects will favor trial balloons, a result we formalize in Theorem 1, though as we show in that result it is still important to consider the settings in which the principal prefers no discovery at all. However, even when a disfavored project has a lower variance, and thus is less effective at improving the agent's beliefs, we show in Theorem 2 that its lower chance of spoiling may still dominate for an intermediate range of favorabilities. These main results also provide two-part conditions which describe the effects of increasing prior precision depending on whether the sender places low or high weight on the disfavored project. We defer full proofs of all results to the relevant appendices, and provide intuitions in the main text.

[5] This choice is important for our results; it reflects a world in which projects are pre-screened to be similar, e.g., a Democratic budget bill generally includes higher spending than a Republican one.

[6] In Section 2.5.1, we also consider the possibility of discovering both values.

[7] In Section 2.5.3, we consider the case where a noisy public signal of v_i is observed, so that v_i is not known with certainty.

[8] In Section 2.4.1, we consider other variations of the principal's utility.

2.4.1 Benchmark Cases

Before characterizing trial balloon optimality, we first introduce some results for simple benchmark cases that restrict the prior distribution of v or the principal's preferences. These insights are useful steps in proving more general results, but may also be of independent interest to model particular settings.

To build intuition, assume $\rho = 1$, so that the two projects are perfectly correlated. This choice abstracts away from the comparison of trial balloons versus building consensus and instead focuses on the simplest trade-off between improving beliefs and spoiling an ex-ante favored project:

Proposition 4 (Perfectly Correlated Projects)**.** *When $\rho = 1$, then discovering one project is the same as discovering both, and the following is optimal:*

1. *If $\mu_1 + \mu_2 \geq 0$: no discovery.*
2. *Else if $\mu_1 < 0$ and $\mu_2 < 0$: full discovery.*
3. *Else $\mu_i = -\mu < 0 \leq c\mu = \mu_j$ for $c \in [0, 1)$. Then there is a cutoff $c^* \in [0, 1]$ such that either*
   (a) *Very disfavored project 1: $c < c^*$, no discovery.*
   (b) *Mildly disfavored project 1: $c \geq c^*$, full discovery.*

Proof. For this proof, and those of other results in this section (propositions 5 and 6), see Appendix B.1. □

Using the structure of the bivariate normal distribution, we obtain a closed-form expression for c^*, including conditions for the cases when $c^* = 0$ so that there is always full discovery or $c^* = 1$ so that there is never full discovery. For small c, the principal is unwilling to risk spoiling the sure thing, but as c grows larger the possibility of improvement becomes more tempting. To showcase the effect of variance, consider the case of $w_1 = w_2 = 1/2$, where we obtain a simple expression, $c^* = 1/(2 - \sigma_1/\sigma_2)$. As overall variance increases (captured by increasing both σ_1 and σ_2), the probabilities of spoiling and improvement both increase. However, even in this simple case the net effect of increased variance is uncertain.

To highlight the role of project variance in improvement, we can modify the principal's preferences so that she *only* gets positive utility if both projects are approved. In this case, straightforward computation of the explicit form of the agent's posterior is all that is needed for the result:

Proposition 5 (Complementary Projects)**.** *Assume the principal only cares about the probability of obtaining the grand bundle, and $\mu_1 + \mu_2 < 0$ so that it cannot be obtained with certainty. When discovering only one project, the higher-variance one is always preferred. Discovering both projects is strictly better than discovering either individually.*

Intuitively, there is no issue of spoiling, because preserving a single ex-ante favored project is useless for a principal who acts as if the projects are perfectly complementary. Thus all that matters is the possibility of an above-mean realization that raises the agent's posterior

belief, which is affected only by the variances of the distributions from which discovered values are drawn. This case captures settings where the principal's agenda control is more limited. For example, a firm advertising an open position can choose which aspects of the job to feature in an ad campaign (and can "commit" to the campaign by rolling out ads before beginning to interview workers) but is only interested in attracting workers who are willing to do all the projects the firm considers part of the job. Our result suggests that firms benefit from focusing on a project where potential employees' opinions are widely dispersed regardless of the average candidate's opinion.

We also consider the case of perfectly substitutable projects, requiring that both projects are ex-ante disfavored to ensure there is some discovery:

Proposition 6 (Substitutable Projects)**.** *Assume the principal cares only about the probability of having at least one project approved. Let $\mu_1 < \mu_2 < 0$ so that approval cannot be obtained with certainty. Discovering the project i satisfying $\sigma_i/\mu_i < \sigma_j/\mu_j$ is better than discovering project j, and discovering both projects is strictly better than discovering either individually.*

As in the previous proposition, there is once again no possibility of spoiling, because the presence of any ex-ante favored project in this environment would trivially lead to no discovery. Thus it is still the ratio of variances that determines which project is discovered, with the modification that a lower-variance project may be discovered if its low mean makes it most likely to yield a positive draw. Full discovery is better than discovery of any individual project because, given both projects have negative means, the law of iterated expectations ensures that $\mathbb{E}[\mathbb{E}[\mu_i|v_j]] = \mu_i < 0$—the average posterior mean of an undiscovered project is unchanged. Thus the principal does better in expectation by discovering that project as well, and chooses to do so by discovering both projects at once if possible. Although this case is less prominent in the real world than that of perfect complementarity, we believe it may still capture some organizational settings. For example, a middle manager may need an executive's approval to hire an additional employee for a team, but only has one position to fill; our result pins down the trade-off between hiring a more favored candidate (as captured by the mean) and hiring one with a better chance of a surprising performance (as captured by the variance).

2.4.2 Optimality of Trial Balloons

We now turn to the general case of $\rho \in (0, 1)$ and a principal with weakly positive weights $w = (w_1, w_2)$ for the two projects satisfying $w_1 + w_2 = 1$. Our first result characterizes optimal single-project discovery when no project is ex-ante open to approval by the agent:

Proposition 7 (Discovery with No Favored Projects)**.** *Let $\mu_1 < \mu_2 < 0$, and let $w_1 = 1/2$. Fixing σ_2, there exists $\underline{\sigma} \in (\sigma_2, (\mu_1/\mu_2)\,\sigma_2)$ such that discovering project 1 is best if and only if $\sigma_1 > \underline{\sigma}$. Otherwise, discovering project 2 is best.*

Proof. For this proof, and those of other results in this section (theorems 1 and 2), as well as some supplementary results, see Appendix B.2. □

Imposing the condition $w_1 = w_2 = 1/2$ on the relative weights allows us to decompose the principal's utility into a "complements term" and a "substitutes term," and invoke the results

of Propositions 5 and 6 to establish which discovery rule maximizes each of those terms. The trial balloon under-performs relative to the complements case but over-performs relative to the substitutes case. Although it is clear that increasing w_i favors discovery of project i, we are unable to verify that the utilities of these discovery rules remain single-crossing. However, thorough simulation[9] suggests that single-crossing is preserved for all weights, suggesting that the relationship above would hold for a decreasing continuous cutoff function $\underline{\sigma}(w_1)$.

To sharpen the distinction between trial balloons and consensus projects, we turn to the case where the former is ex-ante disfavored by the agent, whereas the latter is ex-ante open to approval.

Theorem 1 (Optimal Trial Balloons)**.** *Let $\mu_1 = -\mu < 0 \leq c\mu = \mu_2$ for $c \in [0, 1)$ and let $\sigma_1 \geq \sigma_2$, so that project 1 is a trial balloon.*
For any μ, there is a cutoff $\underline{c}(\mu)$ such that it is optimal to discover neither project when $c < \underline{c}(\mu)$ and to discover project 1 when $c > \underline{c}(\mu)$.
The function $\underline{c}(\mu)$ is continuous and has the following properties:

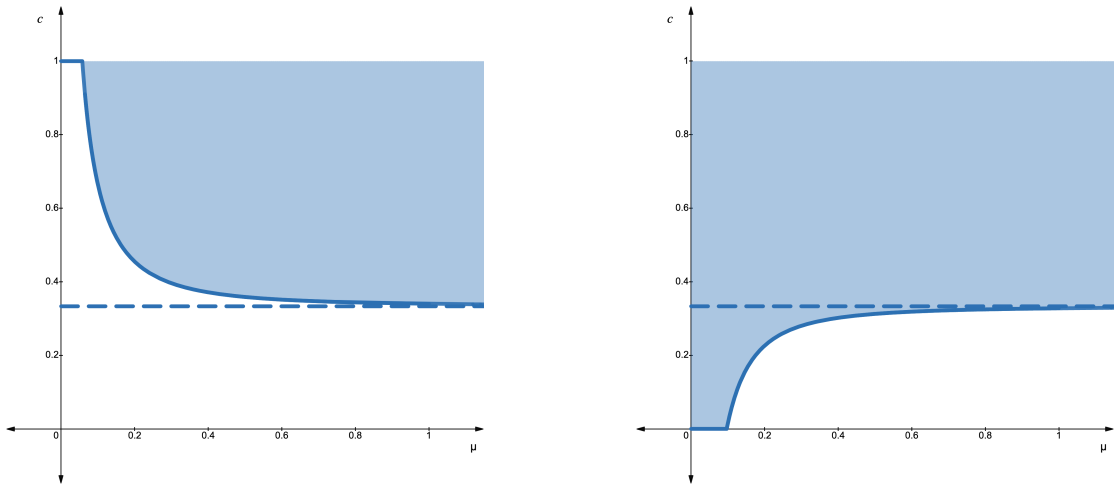
1. *If $w_1 = 1/2$, then*
$$\underline{c}(\mu) = c^* := \frac{\rho\sigma_2}{\sigma_1 + 2\rho\sigma_2} \qquad \text{for all } \mu.$$
2. *If $w_1 > 1/2$, there is $\mu^* \in \mathbb{R}_+$ such that $\underline{c}(\mu) = 0$ for all $\mu \in [0, \mu^*]$. For $\mu > \mu^*$, the cutoff $\underline{c}(\mu)$ is strictly increasing with $\lim_{\mu\to\infty} \underline{c}(\mu) = c^*$.*
3. *If $w_1 < 1/2$, there is $\mu^* \in \mathbb{R}_+$ such that $\underline{c}(\mu) = 1$ for all $\mu \in [0, \mu^*]$. For $\mu > \mu^*$, the cutoff $\underline{c}(\mu)$ is strictly decreasing with $\lim_{\mu\to\infty} \underline{c}(\mu) = c^*$.*

Given the variance condition, the trial balloon is always better than building consensus because it is both less likely to spoil the favored project and more likely to deliver an above-mean realization that attains the grand bundle. The remaining question is whether the trial balloon is better than no discovery at all.[10] A large value of c, meaning that the trial balloon's disfavor is about equal in magnitude to the consensus project's favor, encourages discovery by making spoiling less likely. However, there is a subtle interaction between the minimal such c, the precision of the agent's prior (as captured by taking μ to be large relative to a fixed ratio of σ_1 and σ_2), and the weight w_1 assigned to the trial balloon. When the principal places high importance on the trial balloon, spoiling is not especially concerning compared to obtaining approval for the trial balloon via improvement. Thus, low precision encourages discovery by making both effects strong, and with especially low precision trial balloons are in fact optimal at all values of c. This low cutoff comes with a cost, though: for $c < c^*$, improvement vanishes more quickly than spoiling as precision increases. Thus for a fixed c in this range, no discovery will eventually dominate discovering the trial balloon. Thus the cutoff when $w_1 > 1/2$ must increase in μ. The effect when $w_1 < 1/2$, so the principal places low weight on the trial balloon, is the opposite: the cutoff for trial ballon discovery begins above c^*, so improvement vanishes more slowly than spoiling as μ increases, and for any $c > c^*$ the trial balloon eventually dominates. Figure 2.1 provides an example of the cutoff in each of these two cases.

[9] We use a 1,000 point grid for each of ρ, w_1, μ_1/μ_2, and σ_1/σ_2.

[10] Although we focus on optimality of trial balloons, the proof of this result can also be used to compare discovery of the consensus project to no discovery.

Figure 2.1: Optimal Trial Balloons



An example of Theorem 1: in the left panel $w_1 < 1/2$ and in the right panel $w_1 > 1/2$. The dashed line marks the limit value c^*, and the trial balloon is used when (μ, c) is in the blue shaded region.

In the political context, where we expect priors about a presidential administration's policies to grow more precise later in that president's term, our results make an important testable prediction about the prevalence of trial balloons. If a trial balloon topic is relatively unimportant to the president's priorities, we would expect more trial balloons later in the term; if instead the trial balloon topic is central to the administration's agenda, we would expect early trial balloons followed by convergence to consensus-building proposals made without any discovery. Tracking administration priorities (e.g., through text analysis of press releases) and perceived support for various projects (e.g., through recurring poll questions) would allow us to test this prediction in the data.

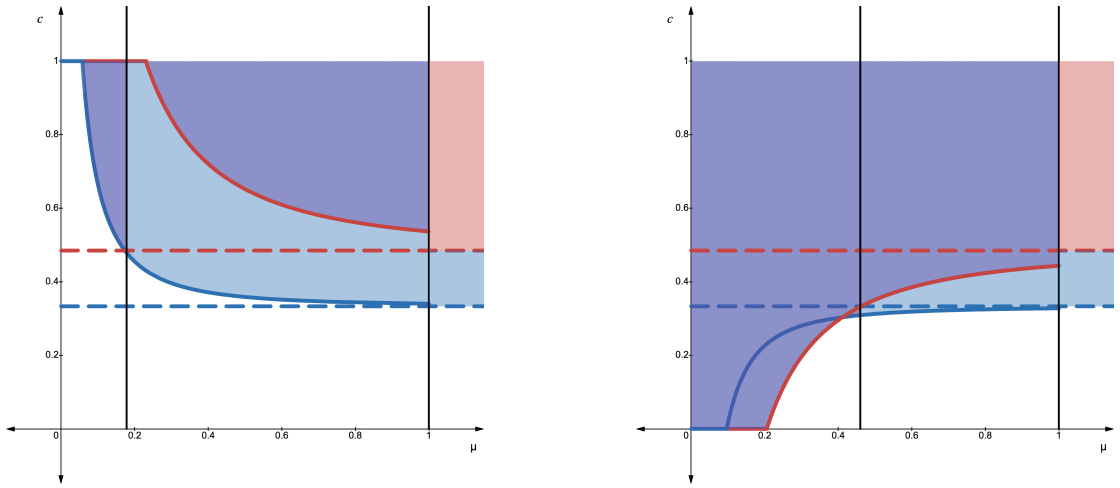
2.4.3 Disfavored Projects with Lower Variance

When $\sigma_1 < \sigma_2$, the disfavored project is no longer a trial balloon; it continues to have a lower risk of spoiling but now has a lower chance of improvement as well. This trade-off alters the choice of optimal discovery rule and carves out room for building consensus. Although the comparison of either project to no discovery is unchanged, comparing the two projects to each other requires careful consideration of the rates at which spoiling and improvement vanish. In the theorem below, we use the limit thresholds at which each project dominates no discovery to make the relevant comparisons.

Theorem 2 (Optimal Single-Project Discovery)**.** *Let $\mu_1 = -\mu < 0 \leq c\mu = \mu_2$ for $c \in [0, 1)$ and let $\sigma_1 < \sigma_2$ so that project 1 is disfavored but not a trial balloon.*
Define $c_i(\mu)$ as the cutoff above which project i is preferred to no discovery. When $w_1 = 1/2$, the cutoffs are constant in μ, with

$$c_2(\mu) = c^{**} := \frac{\sigma_2}{\rho\sigma_1 + 2\sigma_2} > \frac{\rho\sigma_2}{\sigma_1 + 2\rho\sigma_2} =: c^* = c_1(\mu).$$

Figure 2.2: Optimality Regions for Single-Project Discovery



An example of Theorem 2: in the left panel $w_1 < 1/2$ and in the right panel $w_1 > 1/2$. In blue, project 1 is discovered; in red, project 2 is discovered; in purple, an unknown project is discovered. The black vertical lines separate the three cases, with $\mu > 1$ standing in for the limit case.

The value of μ divides optimal discovery into three regions: [11]

1. *Low μ: no discovery if $c < \min\{c_1(\mu), c_2(\mu)\}$ and discovery of an unknown project otherwise.*

2. *High μ: the cutoff $c_1(\mu)$ lies strictly below $c_2(\mu)$. There is no discovery if $c < c_1(\mu)$, discovery of project 1 if $c \in (c, c_1(\mu))$, and discovery of an unknown project if $c > c_1(\mu)$.*

3. *Limit μ: no discovery if $c < c^*$, discovery of project 1 if $c \in (c^*, c^{**})$, and discovery of project 2 if $c > c^{**}$.*

In the limit, the interval $[0, 1)$ of possible values of c can be partitioned into three sub-intervals such that the optimal discovery rule is constant in each sub-interval. However, this property does not hold outside of the limit. Depending on the model parameters, numerical simulations reveal that a variety of patterns for optimal discovery are possible. For example, under the parameters used to produce the right panel of Figure 2,[12] computing the principal's utility shows that the regions where it is optimal to discover project 1 are disconnected.

The clearest conclusion of Theorem 2 is that, for large enough μ, discovering the disfavored project is best for intermediate values of c. Furthermore, if $w_1 < 1/2$ then this intermediate range narrows as precision increases, so that discovery in general becomes more common but disfavored projects as a share of discovered projects become less common. If instead $w_1 > 1/2$, the opposite holds; discovery becomes less common, but more of the discovered

[11] The boundaries of these regions differ depending on the value of w_1. We include a precise characterization in Appendix B.2, and focus here on the behavior of the optimal discovery rule.

[12] Specifically, we set $\sigma_1 = 1/20$, $\sigma_2 = 1/5$, $\rho = 1/4$, and $w_1 = 2/3$.

projects are disfavored.[13] These predictions contrast with those of Theorem 1, and highlight the difference between the trial balloon setting and one in which the consensus project is more effective in improving the agent's beliefs. In the trial-balloon case, the changing range of optimality for the disfavored project depended only on whether spoiling or improvement vanished more quickly. Here, it depends also on the rate at which they vanish compared to the consensus project. When $w_1 < 1/2$, vanishing spoiling enlarges the region where discovering some project is optimal. Because spoiling vanishes more rapidly for the disfavored project, the optimality of discovering it for values ever closer to c^* is overcome by the optimality of building consensus for values ever closer to c^{**}. When instead $w_1 > 1/2$ and vanishing improvement reduces the region where discovering some project is optimal, faster-vanishing improvement for the disfavored project leaves room for the optimal discovery region to grow as $c_2(\mu)$ more slowly approaches c^{**}.

To see the practical implications of the result, consider a politician that chooses not between a trial balloon and a popular low-variance policy, but between a polarized policy and an appealing generality. The polarized policy may be pork-barrel spending that is unpopular on average because only voters in a given district approve; given the definite nature of the spending, there is little chance public opinion will change. The appealing generality may be a promise to cut taxes; the average voter likes the idea, but depending on the details approval will likely change. Suppose further that the politician is self-interested and values a bill containing pork-barrel spending more than one with tax cuts. Then, as the public forms a more precise prior, the politician will be more likely to directly propose legislation without first making public commitments (less discovery overall). However, when the politician does speak in order to gauge the audience's response, they will be more likely to focus on the pork-barrel spending (relatively more discovery of the disfavored project). This prediction draws a sharp distinction between optimal behavior when the disfavored project's variance is lower rather than higher, and could be tested in a similar way to Theorem 1.

2.5 Extensions

The previous section captures the intuitive trade-off between spoiling and improvement that drives the optimality of trial balloons, but makes stark assumptions about the discovery rules available to the principal. In this section, we relax first the assumption of single-dimensional discovery and then that of conclusive evidence to show how our qualitative conclusions are preserved in the face of these changes.

2.5.1 Discovering Both Projects

We begin by considering the possibility of full discovery, where in addition to choosing between no discovery and discovery of a single project, the principal can choose to publicly reveal the value of both projects. For simplicity, we set $w_1 = w_2 = 1/2$, and omit the comparison of single-project discovery rules to no discovery:

[13] Lemma 15 in Appendix B.2 proves this monotonicity property in the limit μ region. However, monotonicity does not hold away from the limit for all parameter values, though it does arise in a high percentage of simulations.

Theorem 3 (Multi-Project Discovery)**.** *Optimal discovery, allowing for discovering both projects, is as follows:*

1. *If $\mu_1 < 0$ and $\mu_2 < 0$, discover both projects.*
2. *Else $\mu_1 = -\mu < 0 \leq c\mu = \mu_2$ for $c \in [0, 1)$. Assume $\sigma_1 \geq \sigma_2$ so that project 1 is a trial balloon. Then, there are values $0 < c_\ell < c_h < 1$ such that*
   (a) *for any $c \in [0, c_\ell)$ there exists $\underline{\mu}$ such that discovering project 1 is better than discovering both projects for all $\mu > \underline{\mu}$.*
   (b) *for any $c \in [c_h, 1)$ there exists $\bar{\mu}$ such that discovering project 1 is worse than discovering both projects for all $\mu > \bar{\mu}$.*

   There is also $c_{no} \in (0, 1)$ such that
   (c) *for any $c \in [0, c_{no})$ there exists a continuous function $\mu_{no}(c)$ such that full discovery is better than no discovery if and only if $\mu \in [0, \mu_{no}(c))$.*
   (d) *for any $c > c_{no}$, full discovery is better than no discovery for all μ.*

Proof. For this proof, and those of other results in this section (propositions 8 and 9), as well as some supplementary results, see Appendix B.3. □

We use numerical simulations similar to those in Proposition 7 to verify that the gap between c_ℓ and c_h is closed at some intermediate c_{mid}, below which case (2a) occurs and above which case (2b) occurs. As with the non-limit cases of Theorem 2, we can find counterexamples for finite μ where the dominance regions for each discovery rule are not connected. Finally, numerical values of c_{no} lie both above and below c_{mid} depending on the other parameters.

Case (1) here parallels Proposition 7 in applying the results of a complements-and-substitutes decomposition of the principal's utility. However, rather than facing a potential trade-off between the two terms which leads to a "compromise" cut-off, in this case the greater variance coming from the ability to discover both projects dominates any single-project discovery rule. Combining case (2) with Theorem 1 leads to a similar intermediate region of dominance for the trial balloon as the disfavored project in Theorem 2. We can also obtain the same limiting comparative statics with respect to increased precision, with increased precision shrinking the dominance region when $w_1 < 1/2$ and expanding it when $w_1 > 1/2$ even as the behavior of the overall discovery region is opposite. Although the limitation to single-project discovery is often natural, this result reinforces the conclusions of Theorem 2 by making clear that depend only on the magnitude and convergence rate of the spoiling and improvement probabilities as functions of some arbitrary discovery rule.

2.5.2 Sequential Discovery

Once the principal is allowed the possibility of discovering both projects at once, it is natural to consider sequential discovery in comparison to the one-shot discovery rule described above. This setting is challenging to work with, but we can establish several basic facts. First, the

principal never discovers both projects at once, and her utility is trivially higher than in the one-shot case. The results of Theorem 1, appropriately modified to use the parameters of the posterior belief after the first discovery step, determine whether a second project is discovered. Finally, the regions where a project will be discovered are larger than in either of Theorems 1 and 3, because the ability to discover a second project offers the principal a chance at "redemption" by discovering a second project afterward.

Two obstacles remain to fully characterize the sequential model: a tight characterization of when the first discovery step is undertaken, and a description of the order of optimal discovery. We leave these for future work.

2.5.3 Noisy Discovery

The conclusive evidence setting may be attractive for some examples, such as an engineer choosing which project to prototype before presenting to management, but it is also a stark restriction on the principal's ability to discover information. In this section we allow the principal to discover a noisy signal of a project's value rather than discovering the value with certainty.

Proposition 8 (Discovery with Exogenous Noise)**.** *Let discovery of issue i generate a noisy signal s_i of its value with distribution $s_i \sim N(v_i, \tau_i^2)$. Then the results of Section 2.4 apply unchanged after substituting the distribution*

$$N\left(\mu_i, \frac{\sigma_i^4}{\tau_i^2 + \sigma_i^2}\right)$$

for the project i which the principal chooses to discover and

$$N\left(\mu_j + \rho\frac{\sigma_j}{\sqrt{\sigma_i^2 + \tau_i^2}}(s_i - \mu_i), (1-\rho)^2\sigma_j^2\right)$$

for the posterior belief about project j, conditional on discovery of project i.

A noisy signal weakens both the spoiling and improvement effects for project i by reducing the amount by which discovery shifts the agent's beliefs. Thus the effect on optimal discovery is uncertain. However, the presence of exogenous noise does not alter the principal's fundamental trade-off; she simply accounts for the signal's noise and considers a problem in the style of Section 2.4 with an appropriately modified prior belief. In general, making the signal of a project's value noisy both makes the principal less likely to discover it and weakly decreases her utility from the optimal discovery rule. This result suggests that future work may benefit from considering the possibility for the principal to reduce the signal's variance at some effort cost.

2.5.4 Discovery with Many Projects

Thus far we have abstracted away from the motivating issue of a complex proposal by considering a case with only two correlated projects, rather than some large but finite number of projects N. This choice is made for tractability, and we can view the two projects in our

model as standing in for the types of projects a principal might have access to in a richer setting. Nevertheless, we can still draw some limited conclusions with an explicit model of $N > 2$ projects.

Let the marginal distribution of the agent's value for project $i \in \{1, 2, ..., N\}$ be $v_i \sim N(\mu_i, \sigma_i^2)$, and assume that the joint distribution of values is an N-variate normal with equal correlation ρ between any pair of projects. Then Propositions 5 and 6, describing perfect complements and perfect substitutes, extend without modification. Specifically, if the principal maximizes the probability of having the grand bundle approved, she always discovers the project with highest variance. If instead she maximizes the probability of having at least one project approved, she discovers any project i satisfying $\sigma_i/\mu_i = \min_{k\in\{1,...,N\}} \sigma_k/\mu_k$. In the latter case, she also always prefers discovering more projects to discovering fewer.

Obtaining parallel results to Theorems 1 and 2 is more challenging; rather than comparing spoiling and improvement effects across two projects, we now have to make the comparison across N projects. These pairwise comparisons are further complicated by the ability to spoil or improve multiple projects at once, rather than only spoiling the single favored project or improving the single disfavored project. To accommodate these issues and preserve our focus on discovery of trial balloons, the following result allows for only one negative-mean project:

Proposition 9 (Optimal Single-Project Discovery with Many Projects)**.**
Let $\mu_1 = -\mu < 0 \leq c_j\mu$ with $c_j \geq 0$ for all projects $j \in \{2, ..., N\}$, and let $\sum_{j\in\{2,...,N\}} c_j \leq 1$. Then the following analogues of previous theorems hold:

- *(Analogue of Theorem 1, part 1): For any project k, choosing one c_j and fixing the others, there is a continuous function $\underline{c}_j^k(\mu, w_1)$ such that discovering that project is better than no discovery if and only if $c_j > \underline{c}_j^k(\mu, w_1)$. This function is decreasing in w_1 for any μ.*

- *(Analogue of Theorem 1, part 2): If $\sigma_1 = \max_{k\in N} \sigma_k$, then discovering project 1 is always better than discovering any other project.*

- *(Analogue of Theorem 2): If $\sigma_i < \max_{k\in N} \sigma_k$, then discovering project 1 has a weaker spoiling effect than discovering project k, but need not have a stronger improvement effect.*

Theorem 1 extends almost unchanged in this setting; the main difference is that we can no longer analytically sign the behavior of the cutoff $\underline{c}_j^k$ with respect to increasing prior precision, though we can still see that the cutoff decreases as more weight is placed on the trial balloon. The expressions for the limit values of each $\underline{c}_j^k$, analogous to c^* or c^{**} in Theorem 2, cannot be written in fully closed form. Thus the conclusions of Theorem 2, which rely on the ordering of those cutoffs, are almost entirely lost; in fact we cannot even be sure that a lower-variance disfavored project does not maintain a stronger improvement effect. These limitations emphasize the complexity of discovery in high-dimensional environments and suggest that a low-dimensional heuristic, such as the spoiling versus improvement trade-off we identify, may serve as a reasonable guide for real-world decisions.

2.6 Conclusion

Real-world policy proposals bundle many correlated projects, and numerous constraints—time, money spent, and attention, to name a few—may restrict a proposer's ability to communicate about the proposal as a whole. To study the choice of which project to emphasize, we present a stylized model with a two-project proposal space where a principal decides whether to publicly discover information about an agent's unknown value in order to gain the agent's approval for their final proposal. In our baseline setting, where the principal can only discover one of the two project values and places additive weight on each project, we characterized the optimality of trial balloons—ex-ante disfavored projects with higher variance than the favored alternative. The optimal discovery rule is driven by a trade-off between spoiling any ex-ante favored projects and improving the agent's beliefs to get more projects approved. We highlight the interplay between the relative weights, magnitudes, and variances of the two projects, which lead to subtle results. Trial balloons, and discovery of any project in general, become more common with increasing prior precision if and only if the weight on a disfavored project is lower than that on a favored one. However, the share of this region taken up by a lower-variance disfavored project has the opposite sign. If it has low importance, then the disfavored project takes up a smaller share of the growing discovery region; with high importance, it takes up a larger share of the shrinking discovery region. Even when the principal has access to richer discovery rules, such as full or noisy discovery, disfavored projects benefit from the ability to explore larger bundles without foreclosing the possibility of a safe proposal if that project fails. Our setting captures only some of the many factors driving the adoption of trial balloons in political and organizational settings, and there remains work to do in tailoring the broad strokes of this framework to particular settings and the discovery rules that fit them.

Chapter 3

Strategic Opinion-Writing on Appellate Courts

Ruling on thousands of cases each year, U.S. federal courts of appeals make some of the most impactful decisions in modern society. I study how polarized partisan environments affect consensus-building using quasi-random three-judge panels on these courts from 1970–2013, documenting a novel pattern in dissenting opinions. Compared to party-unanimous panels, party-mixed panels cause all judges to dissent more often, and at equal rates: median and majority-party judges are just as driven to dissent as their more politically extreme colleagues. This result is incompatible with classical models of judicial politics and is unique to partisanship. To explain my results, I introduce a theoretical framework where judges' favored coalitions are more homogeneous along both partisan and non-partisan dimensions. Using judge metadata, I find suggestive evidence for the model's result that polarization increases dissents by judges of panel-minority law school or gender. With state-of-the-art machine learning tools from natural language processing, I generalize beyond dissents, showing that those same features drive differences in opinion text even when rulings are unanimous. My findings show that partisanship has a powerful and complex effect on agreement and illustrate the need for new tools to capture its consequences in this opaque yet high-stakes environment.

JEL Classification: C45, C79, D72, K10, K40
Keywords: appellate courts, judicial opinions, dissent, natural language processing

Chapter Acknowledgments: I thank Daniel Chen, Wei Lu, Stephen Morris, Benjamin Olken, Jett Pettus, Kathryn Spier, Alexander Wolitzky, and participants at MIT Development/Political Economy lunch for helpful discussions and comments. I also thank the Data Science Justice Collaboratory for providing core appellate court data for this project. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1745302.

3.1 Introduction

Partisan behavior by judges has been spotlighted in the U.S. following recent high-profile, politically impactful decisions by the Supreme Court. Early in his tenure, current Chief Justice John Roberts famously claimed that judges are like baseball umpires who "[only] call balls and strikes," i.e., clarify rules without being political. However, polling from the Supreme Court's most recent term in summer 2024 finds that a majority of Americans believe those justices decide cases based on their personal ideology rather than impartial interpretations of the law [5, 31].

Less remarked upon, but far more influential in terms of volume, is the next level of the U.S. judicial system: thirteen federal courts of appeals which serve as the final word on about 40,000 cases each year. A decades-old empirical literature starting with Sunstein et al. [69] has studied individual judge behavior on these courts using quasi-random assignment of judges to three-judge panels, and of those panels to cases. The interactions *between* judges on panels have been less carefully analyzed, with many prior works treating within-panel decision-making as a black box and focusing directly on the consequences of panel characteristics for outcomes. However, appellate panels provide a rich environment in which to study classical questions from political economy about how partisanship affects consensus-building and coalition formation. In this work, I combine insights from the theory of judicial politics with new tools from computer science to show how partisanship increases overall disagreement among judges, but does not do so along party lines.

To capture the high-stakes consequences of how judges interact on a panel, I focus on *dissents*—the decision to cast a vote against the majority's preferred ruling and the most visible form of disagreement among judges. A key result from prior work is that bipartisan panels—with, e.g., two Democratic judges and one Republican judge—cause a sharp increase in dissenting opinions compared to party-unanimous panels [60]. I first verify this pattern in the full sample of published appellate court opinions from 1970 to 2013, finding a causal increase of about 25% more dissents on bipartisan panels compared to party-unanimous panels. This effect is quantitatively meaningful, accounting for about 1,000 more cases with dissents each year. I also contribute novel evidence that the distance between the panel median and the most extreme judge—not a judge's individual distance to the median—drives this increase.

The importance of the extremal judge suggests that predictions from classical median voter models might explain the increase in dissents on bipartisan panels, i.e., a minority partisan or more polarized extremal judge will dissent more often and raise the total dissent rate. However, I provide robust evidence of a previously unexplored pattern in dissenter ideology: minority partisans are no more likely to dissent than majority partisans. More strikingly, the median judge dissents just as often as the extremal judges; on an average of over 1,300 cases each year, the leftmost and rightmost judges form a majority coalition which excludes the center judge. Thus, judge partisanship causes the *existence* of more dissents, but it does not predict the *identity* of the dissenter. Even taking the bounds of my estimates which are most favorable to partisan differences in dissent rates, 75% of the additional disagreements caused by partisanship do not split the judges along partisan lines.

Combining theoretical and simulated results, I show that these two empirical facts are

incompatible with both a model of sincere dissents with noise and a median-voter-like approach to strategic dissents. I also examine non-partisan judge characteristics—law school training, gender, and age—and show that partisanship is unique among those characteristics in its effect on dissents. The presence of an outlier judge along those non-partisan dimensions is associated with an increase in total dissents; however, that increase is smaller than the one caused by bipartisan panels, and concentrated on the relevant outlier judge. For example, switching a panel where all judges share a law school to one where only two judges share a law school mainly increases dissents by the minority law school judge. Unlike the effect of party, these new results are explained by classical approaches to judicial politics.

To partially address the puzzle of partisan dissents, I describe a framework where judges can differ on both partisan and non-partisan dimensions, and gain more utility from coalitions with more similar judges. The two most similar judges form a potential minimal majority coalition, while the least similar judge will dissent whenever they are so dissimilar that their share of the coalition surplus is less than their outside option. This model provides a link between partisanship and non-partisan characteristics and comes close to explaining the puzzle of partisan dissents. As a simple example, consider a setting where judges' partisan and non-partisan characteristics are orthogonal, so that, e.g., a Republican and a Democrat are equally likely to have gone to Law School X. The increase in dissents on bipartisan panels is then straightforward—party-unanimous panels contain only judges who may or may not share a law school, while bipartisan panels are mechanically more divided along party lines and equally likely to be divided along law school lines. Thus, since reduced similarity increases the chances of dissent, bipartisan panels have a higher overall dissent rate. However, focusing exclusively on bipartisan panels, the full vector of judge characteristics—not just partisanship—determines which judge is least similar and therefore at risk of dissenting. Consider a panel with one Democrat from Law School X, one Democrat from Law School Y, and one Republican from Law School X. Both the partisan and the law school minority judges are tempted to dissent; which one actually does depends on whether partisanship or law school background contributes more to similarity.

The interactions between partisan and non-partisan characteristics are challenging to measure because of reduced sample size. However, I provide suggestive evidence that partisanship activates disagreements along non-partisan dimensions. I compare judges who are outliers in some characteristic to those in the panel majority for that characteristic, and test whether bipartisan panels cause differential increases in those judges' dissent rates. I find a quantitatively meaningful difference in the point-estimated dissent rates of judges in the law school or gender minority on a panel, though statistical power is reduced due to the rarity of shared-law-school and majority-women panels. I then compare these effects to the total increase in dissents caused by partisanship, scaling them using the share of bipartisan panels where they apply. For example, about one-quarter of bipartisan panels have two unique law schools, and on those panels law-school-minority judges dissent 10.9 percentage points more often. If partisanship's only effect was to provoke disagreements along law school lines, then the increase in dissents on bipartisan panels would be about 3%, one-eighth of the observed increase of 23.5%. The combined effects of law school and gender could account for about one-third of observed disagreements on bipartisan panels.

To provide further evidence of latent disagreements along non-partisan dimensions, I use state-of-the-art natural language processing (NLP) tools from the machine learning community

to study opinions even in the absence of observed dissents. While the ruling on a case settles that particular issue, the content of the majority opinion sets legal precedent and decides the consequences of that case for future appellate decisions. An outlier judge who knows majority judges are likely to write a different opinion is thus likely to disagree with the majority. That disagreement may be insufficient to prompt a dissent, or the outlier judge may even strategically join the majority with the hope of swaying opinion content; either of these factors would prevent standard outcome-based measures from detecting disagreement. However, by directly analyzing similarity in opinion text, I can quantify which judge characteristics drive the largest differences in opinion content, and are thus most likely to produce disagreements.

Specifically, I fine-tune a large transformer model [23] to predict legal topics in each opinion, resulting in learned semantic text embeddings: high-dimensional vector representations of opinion text which capture its full semantic context. My first similarity measure follows a recent approach in applied machine learning and uses predicted probabilities as an estimate of the share of opinion text devoted to each topic. I show that opinion text has only minor differences across party; in particular, the topic distribution on bipartisan panels is essentially the same regardless of which party holds the majority. However, the topic distribution is markedly different by gender, complementing recent results which mapped measures of gender bias in opinion text to more conservative rulings on gender-related cases [4].

My second measure of opinion similarity directly leverages the semantic embeddings. Using an approach established in machine learning but largely new to economics, I directly compute the similarities of embedding vectors and look at the neighborhood of each opinion to assess co-location of opinions along characteristics of interest. This approach allows me to directly study the effect of law school, which—since it takes over 150 possible values—could not be easily interpreted as an explanatory variable in previous specifications. Benchmarking against the tendency of opinions to cluster by primary legal topic, I find that neighboring opinions are more likely to be written by judges of the same gender or law school than random chance would predict, while they are if anything slightly *less* likely to be written by judges of the same party than under random chance. Together, these results provide clear evidence that even on panels where a dissent does not occur, non-partisan variables capture a larger and more influential share of legal disagreement than partisan ones.

This work is part of, and contributes to, a growing modern literature studying the effect of political polarization on judicial decision-making. Driven by increased perceptions of court politicization in the modern era, as well as the greater amount of available administrative data on courts and judges, recent work (discussed in detail in Section 3.2) has shown that the effects of partisanship may be subtle, but are detectable and quantitatively impactful. Many of these works have focused on the outcome of an appellate court hearing; for example, which litigants are favored, or whether a lower court's decision is overturned. I augment this analysis by introducing more fine-grained measures of opinion content via additional judge characteristics and the novel use of NLP to study opinion text. Dissents and legal topics are not as immediately related to appellate outcomes, but they play a vital role in shaping the reach and longevity of rulings. Appellate courts are constantly in dialogue with legal precedent, the other levels of the court system, and potential future cases; understanding when and why rulings feature consensus or disagreement is key to understanding whether those rulings are likely to be influential. I also provide a demonstration of the power of modern machine learning tools in political economy. Applications of transformer models and

text embeddings are novel but likely to grow, with large language models (LLMs) making embedding-based measures and their reduced-form predictions or generative output more easily accessible. By studying opinion text even when observable outcomes like dissents or partisan splits are not available, this work shows the power of these tools for extending standard approaches to richer domains.

The remainder of the paper proceeds as follows: I begin by describing related literature on judicial partisanship in economics and law, as well as applications of NLP to legal domains both within and beyond economics, in Section 3.2. Sections 3.3 and 3.4 describe key institutional details, features of the appellate court data, and my empirical approach. Section 3.5 applies classical models of judicial politics to my setting and introduces my coalition formation framework, providing key testable implications of both. I then present evidence for my two central empirical facts—partisanship causes dissents, and politically-extreme judges do not dissent more often—in Section 3.6. In Section 3.7, I show that this pattern is unique to partisanship and provide suggestive evidence that disagreements along non-partisan dimensions are activated on bipartisan panels, increasing dissents by outlier judges. Section 3.8 supplements that analysis with NLP measures that apply even to unanimous decisions, showing that non-partisan characteristics have more influence on opinion content than partisan ones. Finally, Section 3.9 concludes with a discussion of possibilities for future research, both on the methodological side and on real-world questions around judicial politics. Appendices C.1–C.6 contain additional visualizations and robustness checks for the data and results in this work; each begins with a short paragraph describing its contents in more detail.

3.2 Related Literature

My work contributes to the modern literature using quasi-random assignment to study the effect of appellate judge and panel characteristics on outcomes. This literature includes work using opinions hand-coded as liberal or conservative [49, 69], number of Democratic-appointed judges [29, 30], and more modern machine learning tools [60] to measure panel and outcome partisanship. In general, these works find that judge or panel partisanship causes different outcomes, such as partisan-aligned opinions and reversals. Rather than comparing Democratic- and Republican-dominated environments, this work focuses instead on the difference between party-unanimous and bipartisan ones.

Closely related to this work's analysis of partisanship are Lu and Chen [60] and Chen et al. [26]. The former uses machine learning tools similar to those in Section 3.8 to construct an opinion polarization measure based on prediction accuracy. I instead study legal topics and use both predictions and the full semantic embeddings of opinions. As an auxiliary result, Lu and Chen [60] documents an increase of about one percentage point in the dissent rate of partisan-minority judges compared to partisan-majority judges. This effect does not appear in my sample, likely due to small differences in data cleaning, and is not sufficient to account for the large increase in total dissents on bipartisan panels.[1] I also use judge ideology scores to show that the median judge contributes approximately one-third of the dissenting votes in

[1] I document the full data preparation and provide code to produce all results and tables in this work as part of the replication archive available on my website.

my sample, which would remain a puzzle for standard models even if minority partisans were more likely to dissent.

The latter work, [26], studies a judge's individual distance from the panel and circuit median. The authors document a U-shaped pattern of dissents relative to the panel median and an M-shaped pattern relative to the circuit median, which they explain with a model where judges trade off their personal disagreement with an opinion against social disapproval for dissenting. I complement this across-panel explanation with richer within-panel analysis. In Section 3.6 I show that the most extreme judge on a panel is responsible for increased dissents distributed evenly across judges. Using the model of Section 3.5.3 and the data in Section 3.7.3, I provide suggestive evidence of how partisanship and other judge characteristics may interact to drive dissents.

Turning to non-partisan judge characteristics, prior work has studied the causal effect of race [50] and gender [4] on rulings in the subset of cases where those features are particularly salient. I contribute to this literature by providing a general theoretical framework for similarity between judges, encompassing partisanship, demographics, and other characteristics such as law school training. Empirically, the success of my NLP approach to topic modeling introduces the possibility of extending these works by testing whether the effects of judge identity generalizs to cases where that identity is less salient.

The theoretical literature on appellate decisions has largely elided any separation between partisan and non-partisan judge characteristics by focusing on a unidimensional policy space [22, 65]. Predictions about the degree to which the median judge controls opinion content vary, but a general result is that less polarized judges are more likely to vote strategically and thus less likely to dissent. In my sample, the substantial share of dissents by the median judge suggests that a richer policy space or different preferences over coalitions may be important components of strategic behavior. The former approach is investigated empirically in Lauderdale and Clark [56], which uses word-level statistical data from the U.S. Supreme Court to show that the identity of the median justice varies depending on which topics are most important to a given case. The latter approach is the focus of my framework in Section 3.5.3, which directly considers judge preferences over possible majority coalitions and describes both the settings in which dissents are more frequent and the relevant traits of dissenting judges.

Looking beyond this work's focus on partisanship and dissenting opinions, text data and NLP have seen growing applications in economics [3], political science [67], and law [51]. My work uses NLP tools as a supplement to extensive metadata on cases, judges, and panels. Given the rarity of observable features like dissents, as well as the limited number of hand-coded case topics and outcomes, novel applications of these tools are vital for generalizing existing conclusions about judge behavior and are often the only way to measure properties like opinion similarity at scale. Adapting existing machine learning approaches for the judicial setting, developing new NLP tools to tackle problems of real-world interest, and ensuring that those tools are compatible with existing frameworks for causal inference and estimation, are important technical challenges which I hope to address in future work.

3.3 Setting

3.3.1 Structure of US Appellate Courts

There are thirteen circuit courts of appeals. Eleven cover geographically contiguous sets of states, jointly containing all fifty; they hear appeals from one or more district courts (the next level down in the federal court system) located in their area. There is also a circuit court covering Washington, D.C. only and a Federal Circuit which hears cases not tied to a particular district.[2] The total modern caseload across all circuit courts combined is about 40,000 cases per year; larger circuits, like the Ninth, hear more cases than small ones, like the Seventh. About 7,500 of these cases are filed for further appeal with the Supreme Court, but less than 200 of those are actually reviewed. Thus, for about 99.5% of cases, a circuit court provides the final word.

Circuit court judges are appointed by the president, with approval historically requiring a two-thirds majority vote in the Senate.[3] Seats on the court are assigned to particular districts; each judge will have a "duty station" located in a given district, and the court will convene there for part of the year to hear appeals. The senators for the state containing that district, called "home-state senators," have significant influence on the appointment of the corresponding appellate judge [41]. This influence is reflected in some commonly-used measures of judicial ideology, which take into account the political leanings of home-state senators (see Appendix C.1.2 for details).

3.3.2 The Opinion-Writing Process

Appellate courts decide whether to uphold or reverse a lower court's ruling; they do not make independent determinations of, e.g., guilt or innocence. Their cases cover a broad range of topics—criminal sentencing, constitutional law, antitrust, etc.—and the vast majority are heard by a panel of three judges chosen quasi-randomly (see Section 3.3.3 for a thorough discussion). A select subset of cases deemed particularly challenging are heard by the entire circuit, which is referred to as an *en banc* case. As is common in work on appellate courts, my sample omits *en banc* cases.

Some cases, which are routine or have procedural issues, are decided without oral arguments after reading the submitted legal briefs. My work, and almost all other work outside law itself, considers cases where appellate judges hear oral arguments and discuss them in order to decide on a *disposition*: whether to uphold or reverse a lower court's ruling. If a judge votes against the majority disposition, they are *dissenting*, and write a separate non-binding opinion. If a judge agrees with the majority disposition but disagrees with the reasoning, they are *concurring*, and often write a separate non-binding opinion. The senior judge among those who agree on both disposition and reasoning assigns the writing of the *lead opinion* to a member of that group.[4] This assignment is opaque and may be strategic, driven by

[2] See Figure C.4 in Appendix C.1 for a visualization of circuits and districts.

[3] The threshold is now a simple majority following procedural changes during President Obama's second term. This work's data ends before that change was made.

[4] The term "lead opinion" accounts for the possibility that only one judge remains in the group, i.e., one judge dissents and the remaining two have different reasoning or all three judges agree but with different

personal considerations, or affected by availability and other bureaucratic reasons.

In cases where an opinion sets legal precedent, its text is published; unpublished opinions are more routine and feature a lower share of dissents and concurrences. My sample contains only published opinions, which are generally believed to be more affected by political and strategic behavior [29].

3.3.3 Quasi-Random Assignment on Appellate Courts

Quasi-random assignment of appellate judges to three-judge panels has been treated as an established fact in social science for decades. However, the details of the assignment process are complex. Several law review publications, in particular Levy and Chilton [59], provide a thorough discussion of the history and current practice of appellate panel formations; I summarize key facts here.

Circuits make schedules in advance for several months' worth of cases, and often for the entire year. There is no centralized scheduling procedure; each circuit has its own guidelines, procedures, and institutional norms. Assignment of judges to panels, and panels to cases, may be the responsibility of the chief judge, the clerk of court, the circuit executive, or another member of the court staff; it may be done by computer (e.g., shuffling an Excel spreadsheet), by hand (e.g., drawing names from a hat), or both. The Fourth, Fifth, Sixth, Seventh, Ninth, Tenth, and Eleventh Circuits explicitly state in official procedure documents that they randomly assign judges to panels. Several judges, lawyers, and legal scholars have stated that panels are randomly drawn on other circuits as well [59].

There are several reasons why panel assignment may not be truly random despite these statements. First, some circuits explicitly note that they attempt to implement what amounts to stratified random assignment. For example, the Seventh Circuit states that panels are adjusted so that judges sit "approximately the same number of times with every other judge of the court" over a two-year period [59]. Additionally, across all circuits schedules may be adjusted in advance so that, e.g., a judge is not forced to hear cases for two months straight despite their name appearing in all of the first 50 panels drawn. Finally, schedules are also necessarily adjusted after being finalized, e.g., if a judge is ill or if a case is withdrawn. However, the consensus among both legal scholars and practitioners is that these deviations from randomness are orthogonal to my variables of interest: case topics, ideology, and most judge demographics.

Due to the idiosyncrasies of scheduling, testing random assignment by compiling a full calendar of panels and the cases they heard is a monumental task even in modern data. Instead, two practical tests are used in prior literature. The first is to perform balance checks on observed panels—e.g., in the observed sample of cases from the Ninth Circuit in 1975, do panels of three Republicans hear First Amendment cases about as often as panels of three Democrats (assuming both types of panel are equally common)? The second is to simulate panels and check whether the observed distribution of panel characteristics is an outlier in the simulation—e.g., using the actual judges serving on the Ninth Circuit in

reasoning. In these cases, the senior judge's opinion is the lead and other judges who agree with the disposition are considered concurring. Other than these two (rare) special cases, the colloquial label of "majority opinion" is equivalent to that of lead opinion.

1975, randomly generate the observed number of panels and count the share which have two Democratic judges and one Republican judge; repeat this simulated circuit × year 1,000 times and compare the distribution of shares to the observed share. I present results from both methods alongside detailed descriptions of the exact procedures used in Appendix C.2.

In the appellate court data used here, the observed distributions of case topics, judge characteristics, and panel characteristics are extremely close to what is observed in the "truly random" simulated panels. Depending on the method used, some statistically significant deviations from randomness are observed, but the magnitude of those deviations is generally smaller than half a percentage point. This result accords with the recent literature studying political behavior by appellate judges.[5]

3.4 Data and Empirical Strategy

3.4.1 Opinion Data

My main data is provided by the Data Science Justice Collaboratory; more details are in Ash et al. [4]. Text is originally compiled by Bloomberg Law and contains the universe of published opinions from the 11 geographic circuits, the DC Circuit, and the Federal Circuit from 1890–2013. Each opinion is associated with the following metadata:

- The circuit and year where the case producing the opinion was heard.
- Whether the opinion upheld or reversed the lower court's ruling.
- The names of the three judges on the panel producing the opinion.
- The votes of the three judges (i.e., whether each judge supported the lead opinion, dissented, or concurred).
- The author(s) of the lead, dissenting and concurring opinions (if any).
- The text(s) of the lead, dissenting and concurring opinions (if any).
- Case and opinion metadata; of particular interest is a hand-coded primary topic from Bloomberg Law in one of eight categories: criminal, civil rights, First Amendment, due process, privacy, labor relations, economic activity and regulation, miscellaneous.

In Table C.1, I present summary statistics for opinion text; full distributions of relevant variables are in Appendix C.1.1. Both dissents and concurrences are relatively rare. For opinions where the lead/majority has a known author and contains some text, 9.9% have a dissent with a known author; 99.4% of those dissents have some text, and 73.4% have at least one single-spaced page (500 words) of text. Again restricting to opinions where the lead/majority has a known author and contains some text, 6.8% have a concurrence with a known author; 74.9% of those concurrences have some text, and 38.7% have at least one single-spaced page of text.

[5]For further results confirming random assignment in similar data, see Cohen [29], Ash et al. [4], and Chen and Yeh [25].

3.4.2 Judge Data

Judge characteristics are originally from Appeals Court Attribute Data, the Federal Judicial Center, data collection by Chen and Yeh [25], and Bonica [15]. They include age, gender, region of birth, law school attended, and prior experience. Full distributions of relevant variables are in Appendix C.1.2.

Critical to my analysis are measures of judge ideology, which is challenging to quantify because appellate judges are not elected and vote in panels too small for structural estimation methods like DW-Nominate [66] to apply. Still, I can rely on three existing measures. The first is self-reported party, which is almost always the same as party of appointing president, but may have slight variation over time. The second is Judicial Common Space (JCS) scores from Epstein et al. [35], which are based on the ideology of senators involved in judicial appointment. JCS scores are the gold standard measure of judicial ideology and are the most commonly used measure across economics, political science, and law. The third measure is Database on Ideology, Money in Politics, and Elections—Campaign Finance (DIME-CF) scores, from Bonica and Sen [16], which measure judge ideology using campaign contributions. In many cases JCS and DIME-CF scores are highly correlated, but the authors of the latter claim it captures preferences over a broader set of issues. More details of these ideology measures are in Appendix C.1.3.

While party affiliation is available for over 99% of judges, ideology scores are generally only available for judges sitting on panels from 1970–2013. This time period also coincides with a period of greater stability in judges' partisan stances following the passage of major civil rights legislation in the 1960s; as such, I focus primarily on the shorter time period and supplement my analysis with the full range of panels when possible.

3.4.3 Learned Semantic Embeddings

To quantify properties of opinion text, I follow the state-of-the-art approach in natural language processing (NLP) and use machine learning to generate *embeddings*: high-dimensional vectors which capture the full semantic content of the text and are then used for downstream prediction, classification, and clustering tasks. These embeddings can be generated off-the-shelf from existing large transformer models such as BERT [32], or from large language models (LLMs) such as GPT-4o from OpenAI and LLaMA from Meta. However, their performance for particular tasks can be improved by fine-tuning those models on a supervised learning task. In Appendix C.4, I describe a typical fine-tuning pipeline as deployed in this work, which I use to generate the embeddings used in Section 3.8. I also present performance metrics for the main model used (tables C.15 and C.16 in Appendix C.4), which accurately classifies approximately 75% of opinions into the eight Bloomberg Law topics described in Section 3.4.1.

Text embeddings can be used directly as numerical representations of qualitative data. For example, I can compute measures of similarity between opinions by computing distances between embeddings, or assess clustering of opinions by plotting them (after reducing their dimension using an appropriate algorithm). They can also be used to generate reduced-form data, such as predictions of author party or topic distributions. The latter is more common in applied work (e.g., Lu and Chen 60). In Section 3.8, I work with both reduced-form

predictions and full embedding vectors, and discuss my approach in more detail there.

3.4.4 Empirical Strategy: Quasi-Random Assignment

To identify the causal effect of partisanship on dissents, I leverage quasi-random assignment of judges within each circuit and year to three-judge panels. If assignment of judges to panels were truly random, then I could use specifications of the form

$$\text{Outcome}_p = \alpha \, \text{panel type}_p + \gamma_{ct} + \varepsilon_{pct}, \tag{3.1}$$

where each observation is a panel p on circuit c in year t. The circuit $\times$ year fixed effects γ_{ct} ensure that the panel type on the left-hand side (e.g., whether the panel is bipartisan or party-unanimous) is randomly assigned and is not influenced by omitted variables (e.g., the case topic) or reverse causality from the outcome. They also account for the varying composition of each circuit from year to year, which affects the probability of each panel type, e.g., there may be a greater share of Democrats on the First Circuit in 1975 than on the Seventh Circuit in 2000, and thus the chances of a panel with 3 Democrats is higher on the former than the latter. After including fixed effects, the estimated coefficient $\hat{\alpha}$ represents the causal effect on outcomes of a change in panel characteristics.

As discussed in Section 3.3.3, assignment is not strictly random due to institutional features, but is quasi-random with respect to characteristics of interest in my sample. Still, I generally include a control for legal topic to increase the precision of my estimates and sometimes include judge fixed effects, controls for panel characteristics (e.g., number of unique law schools represented), or controls for author characteristics.

3.5 Modeling Dissents on Bipartisan Panels

Before presenting my empirical results, I describe results from a model of sincere dissents with noise and show that it cannot plausibly explain the equal dissent rates of median and extremal judges. In Appendix C.3.2, I enrich this model by allowing judges to make strategic choices considering both case dispositions and opinion locations in policy space, and show that the structure of the cost of dissent determines the effect of increased distance to the mean on dissent rates by different judges (see Proposition 11 in Appendix C.3.2 and the discussion at the end of that appendix). However, even in this richer model, the inability to fully describe empirically observed patterns in dissents remains.

As a partial explanation of those equal dissent rates, I propose a framework where judges are multidimensional and have preferences over coalitions rather than opinions. This approach comes closer to explaining the patterns in partisan dissents, but my ability to test its predictions about the effect of non-partisan judge characteristics is limited by sample size issues (see Section 3.7). However, it provides both a direction for theoretical progress and important guidance as to what data may be most useful for reconciling the puzzle of partisan dissents.

3.5.1 Sincere Dissents with Noise

I begin by presenting a simple model in the style of Carrubba et al. [22] where judges vote sincerely on whether to uphold or reverse a lower court's decision. Consider three judges $i \in \{1, 2, 3\}$, who have cutoffs $x_i \in [0, 1]$ below which they would like to uphold a ruling and above which they prefer to overturn it. Without loss of generality, let $x_1 < x_2 < x_3$, and let $x_2 - x_1 < x_2 - x_3$, so that judge 2 is the median and judge 1 is closer to the median than judge 3. I will refer to judge 3 as "extremal" and judge 2 as "second-extremal."

A case $c \in [0, 1]$ is a value drawn from some distribution $F(c)$; for simplicity, let $F(c) \sim U[0, 1]$. Then, under sincere dispositional voting, judge 1 dissents if $c \in (x_1, x_2)$ and judge 3 dissents if $c \in (x_2, x_3)$; judge 2 never dissents. By uniformity of $F(c)$, judge 3 dissents more than judge 2, and increasing either judge's distance from the median by one unit increases their probability of dissent by one unit.

Adding case-specific noise is analytically complex, even when the noise is mean-zero noise with a uniform or normal distribution. However, if noise is sufficiently small (e.g., i.i.d. uniform with sufficiently small variance), I can show that increasing the extremal judge's distance has more of an effect on total dissents than increasing the second-extremal judge's distance (Proposition 10 in Appendix C.3.1). Intuitively, noise may create any of the six possible permutations in judge order, but with sufficiently small noise I can focus only on three: the original ordering, the order $x_2 < x_1 < x_3$ with judge 1 as the new median, and the order $x_1 < x_3 < x_2$ with judge 3 as the new median. In the first permutation, increasing judge 1 or judge 3's and distance from judge 2 has the same positive effect on total dissents. However, in the second permutation, only increasing judge 3's distance affects total dissents; in the third permutation, only increasing judge 1's distance affects total disssents. The second permutation is more likely than the third, so the dampening effect caused by permuted judge orders will be stronger for second-extremal distance than for extremal distance.

3.5.2 Simulating Sincere Dissents

Noise produces two attractive conclusions which, as I will show in Section 3.6, are present in the data: the median judge's dissent rate is greater than zero (see Table C.3), and increasing extremal distance has a stronger effect than increasing second-extremal distance (see Table C.2). However, simulations show that neither effect is sufficiently strong to account for the observed results, and a better fit on one implies a worse fit on the other. To conduct this simulation, I define a grid of variances σ^2 between 0 and 0.5; I choose this range because the sample variance of JCS ideology scores is 0.158 (Figure C.11, Appendix C.1.3). For each variance, I draw 100,000 random panels from the data and add i.i.d. $N(0, \sigma^2)$ noise to each JCS score, then compute the share of post-noise panels for which the original median, second-extremal, and extremal judges are now the post-noise median judge. Figure C.1 shows the results.

The median judge's share of dissents is upper-bounded by the share of permuted panels where they are no longer the median; however, reaching this upper bound requires that only the median judge dissents on those panels. A more reasonable assumption is that the median judge and the remaining non-permuted judge are equally likely to dissent, and thus the median judge accounts for half of those dissents. In Section 3.6.2, I show that the

upper bound of a 95% confidence interval for the difference in dissent rates across judges produces a median dissent share 9% lower than that of non-median judges (27% compared to 36%).[6] This rate thus requires at least 27% of panels where the median judge is permuted to another role, and realistically requires about 54% of panels. The left image of Figure C.1 shows that the latter is achieved when the noise has variance approximately equal to the sample variance of JCS scores. Thus, even using the most model-compatible difference in dissents, the required variance makes noise just as informative about judge partisanship as well-documented, historically predictive ideology scores.[7] Since the growth in permuted panels is approximately logarithmic, even increasing noise variance to triple that of JCS scores only delivers a median dissent share of 30.5%, still 2.7% short of the point estimate in Table C.3 (Section 3.6.2).

The greater effect of increased extremal distance compared to increased second-extremal distance comes from the lower probability of permutations making the original extremal judge into the post-noise median. However, for σ^2 large enough, the difference in the probabilities of becoming the post-noise median judge for the original extremal and second-extremal judges must decrease.[8] The right image of Figure C.1 shows that in my simulations, the difference monotonically decreases in σ^2 for all $\sigma^2 > 0.03$. For $\sigma^2 = 0.158$, the value needed to fit the 95% confidence interval of median dissent shares, the difference is about 15%. In Table C.2 (Section 3.6.1), I find that the lower bound of the 95% confidence interval for the difference between the effect of increased extremal distance and that of increased second-extremal distance is about 25%, best fit by the difference-maximizing variance $\sigma^2 = 0.03$. Thus, this fact is even more challenging to fit, and the best approximation dramatically reduces the fit with median dissent rates.

3.5.3 A Framework for Coalition Formation

I now step back from the unidimensional model of cases and propose first steps towards a model of multidimensional judges who directly have preferences over the majority coalitions that can be formed. Full details of this framework are in Appendix C.3.3; I work with a simple example here for ease of exposition. Let judges be two-dimensional, with a type $t_i = (x_i, y_i)$ consisting of their party $x_i \in \{D, R\}$ and law school $y_i \in \{0, 1\}$. The value of a majority coalition is given by the average weighted similarity of its members; two judges who share a party get similarity weight w_x while two judges who share a law school get similarity weight w_y, and the overall similarity is the sum of those weights. Additionally, unanimous coalitions are scaled by $\alpha > 1$, representing the additional value of unanimity even if judges are more dissimilar, and dissenting judges receive some payoff $d \geq 0$. A majority coalition is stable if its value can be divided among member judges so that no subset of judges can form a more valuable coalition.

[6] Of course, about half of the confidence interval is composed of median dissent shares *higher* than those of non-median judges, which this model can never explain.

[7] In Section 3.6.1, I provide strong evidence that JCS scores are predictive of dissents, especially on bipartisan panels, making it even less likely that they are weakly less informative about judge preferences than an arbitrary noise term.

[8] Proposition 10 (Appendix C.3.1) shows that the difference is strictly positive for some σ^2, but it is necessarily 0 for arbitrarily large σ^2, where the original ideology score is irrelevant.

Consider a panel with Democratic judges and judges from Law School Zero in the majority. If there is only one dimension of difference, then the two majority judges may want to exclude the dissimilar minority judge from the unanimous opinion.[9] More interesting is the case where there is a 2-1 split along both dimensions. Then there are two possibilities:

1. With probability 1/3, there is one double-minority judge with type $(R, 1)$.

2. With probability 2/3, the lone R judge has characteristic type 0.

In the first case, the unanimous opinion is stable if and only if it beats the 2-1 split with both identical judges in the majority and the "double-minority" $(R, 1)$ judge dissenting. In the second case, with $w_x = w_y$, the 2-1 splits with the $(R, 0)$ and $(D, 1)$ judges in the minority deliver equal value to their members. (The majority coalition with the $(D, 0)$ judge in the minority may be stable for low d and α, but can be ruled out by increasing either of those parameters.) If instead $w_x < w_y$, the 2-1 split with the $(D, 1)$ judge in the minority will deliver higher value and thus be the main challenge to the unanimous opinion.

This simple setup can largely account for the two key empirical facts of Section 3.6. First, bipartisan panels will have more dissents than party-unanimous panels, since on a party-unanimous panel there is only one possible dimension of disagreement while on a bipartisan panel there are two. Second, if $w_x << w_y$, then I can approximate the observed result that the minority partisan makes up a 1/3 share of dissents. Under that assumption, the $(R, 1)$ judge always dissents when present; otherwise, the $(D, 1)$ judge dissents much more often than the $(R, 0)$ judge. Since the $(D, 1)$ judge is present on 2/3 of panels and the $(R, 1)$ judge is present on only 1/3 of panels, slightly less than 2/3 of dissents are attributed to D judges and slightly more than 1/3 to R judges. As w_x increases to equal w_y, the share of R dissents judge increases to 1/2. However, this result can be offset by adding party-correlated noise to the partisan component of judge types. As seen in the sincere dissents model, this noise reduces dissents by R judges on net by occasionally making a D judge the post-noise partisan minority.

This framework is only a first step towards explaining judges' preferences for unanimous or split decisions. Placing more structure on preferences at the judge level and explicitly specifying a choice rule among stable opinions would allow me to directly estimate a model of this style in the data and compare its accuracy to the noisy sincere dissents approach in the previous setting. However, these first results already suggest an important direction for empirical analysis: measuring the interaction between partisanship and non-partisan outlier status. In Section 3.7.3, I do so (to the extent feasible given available data) and find suggestive evidence supporting two predictions: double-minority judges are most likely to dissent, and characteristic-minority judges dissent more than minority partisans when both are present. Before turning to the joint analysis of party and non-partisan characteristics, though, in the next section I first robustly establish the effect of partisanship alone on dissents.

[9] There is insufficient data to test that, e.g., bipartisan panels with only men from the same law school have higher dissents by minority partisans than majority partisans. In the sample of men-only panels, the minority partisan is no more likely to dissent, but this sample is large and includes many other potential dimensions of judge heterogeneity.

3.6 Relating Partisanship and Dissents

3.6.1 Partisanship Causes More Dissents

Prior literature has found that judges' votes on an appellate panel can be partially attributed to partisanship—both of an individual judge, and of the panel as a whole [29, 49, 69]. In this section I establish the degree to which partisanship is causally relevant for dissents in my sample, which contains all published appellate opinions from 1970 until 2013.[10] My first specification, at the panel level, follows existing approaches and compares bipartisan panels to party-unanimous ones under the setup of Equation (3.1):

$$\text{dissent}_p = \alpha\, \text{bipartisan}_p + \text{controls}_p + \gamma_{ct} + \varepsilon_{pct}. \tag{3.2}$$

"dissent" on the left-hand side is an indicator for the presence of a dissenting vote on panel p (observed on circuit c in year t). On the right-hand side, "bipartisan" is an indicator for whether panel p has two judges of one party and one of the other;[11] γ_{ct} is the circuit $\times$ year fixed effect ensuring quasi-random assignment. To correct for minor deviations from exact balance across judge and case characteristics, in this estimation and all others in this section I report results controlling for potentially relevant panel-level variables: number of unique law schools and number of women on the panel, as well as the Bloomberg Law topic of the case. (Results are unchanged with and without these controls.) I estimate the model as a conditional logit (since the outcome variable is a probability of dissent, and the fixed effects stratify the data) and cluster robust standard errors at the circuit $\times$ year level. The estimate $\hat{\alpha}$, the causal effect of switching from a party-unanimous panel to a bipartisan one, appears in the first column of Table C.2.

Adding a minority-party member to a party-unanimous panel causes a 23.5% increase in dissents, equal to 2.0 percentage points (p.p.), i.e., from 8.4% to 10.4%, or roughly 800 more cases with dissents each year under the modern appellate caseload.[12] This effect is in line with estimates in earlier literature.

My second specification allows for finer variation in judges' political leanings by using ideology scores rather than the bipartisan indicator:

$$\begin{aligned}\text{dissent}_p =& \beta_1\, \text{extremal-dist}_p + \beta_2\, \text{2nd-extremal-dist}_p + \beta_3\, |\text{median}|_p \\ &+ \text{controls}_p + \gamma_{ct} + \varepsilon_{pct}.\end{aligned} \tag{3.3}$$

I again estimate the specification as a conditional logit and cluster robust standard errors at the circuit $\times$ year level. The new variables of interest are the three distance measures, which replace the bipartisan panel indicator in Equation (3.2). For each panel, I compute the median

[10] As discussed in Section 3.4.2, when using ideology scores I focus on the time period 1970–2013, where those scores are nearly complete. All results using party only (and thus the full sample, 1890–2013) are robust to restricting the sample to include only 1970–2013.

[11] Results are robust to separating Democratic- and Republican-majority panels; see tables C.17 and C.18 in Appendix C.5.1.

[12] If the single indicator for bipartisan panels in Equation (3.2) is replaced with two separate indicators for Democratic-majority bipartisan panels and Republican-majority bipartisan panels, the respective coefficients are 0.255 (0.026) and 0.223 (0.024), both significantly different from 0 at a 1% level and not significantly different from each other at a 10% level.

score, and label the judge who is furthest from the median as extremal (the second-furthest judge is thus second-extremal). Both of those variables appear on the right-hand side, as does the location of the median.[13] I also include panel-level controls as in Equation (3.2). The estimated coefficient $\hat{\beta}_1$ then measures the causal effect of moving the extremal judge one unit further from the median judge (holding fixed second-extremal distance and median location), while $\hat{\beta}_2$ measures the effect of moving the second-extremal judge one unit further from the median judge (holding fixed an extremal judge who remains further from the median). The results are in columns (2)–(7) of Table C.2.

Focusing first on columns (2)–(4), which use JCS scores, $\hat{\beta}_1$ corresponds to a 24.1% increase in dissents from a one-standard-deviation increase in the extremal judge's score. $\hat{\beta}_2$ is both much smaller—only a 3.2% increase—and not significantly different from 0. This result is driven by bipartisan panels; changing either distance has no effect on dissent rates of party-unanimous panels. This result matches the intuition that JCS scores capture precisely the legal settings most aligned with left-right ideology. The increased influence of the extremal judge is less clear when using DIME-CF scores, seen in columns (5)–(7). In the full sample, $\hat{\beta}_1$ correesponds to a 13.6% increase in dissents from a one-standard-deviation increase in extremal judge polarization, while $\hat{\beta}_2$ corresponds to a statistically significant 9.7% increase in dissents from the same increase in second-extremal judge polarization. It is once again true that the effect is driven by bipartisan panels, but on that sub-sample $\hat{\beta}_1$ is only significant at the 10% level and is approximately half the size of $\hat{\beta}_2$. The latter result suggests that DIME-CF scores may capture a broader cross-section of legal ideology and thus be a noisier measure on bipartisan panels than JCS scores.

These results provide some support for the median-voter-like predictions of the model in Section 3.5.1 (and the extension in Appendix C.3.2), where the presence of noise (or strategic considerations) could lead the second-extremal judge to have a lesser effect on total dissents than the extremal judge. However, that approach predicts that the median judge will dissent markedly less than their colleagues, even in the presence of very large noise (as shown in Section 3.5.2). To directly test this conclusion, and explore more general departures from classical median voter frameworks, in the next section I separate out dissents according to the "ideological role"—median, second-extremal, or extremal—of the dissenting judge. This approach clarifies whether the increased distance of a judge from the median affects only their own probability of dissent or also the behavior of their fellow panel members.

3.6.2 Minority Partisans Are Not Dissenting More

To show that partisanship fails to predict *who* dissents, I replace the binary dissent indicator on the left-hand side of equations (3.2) and (3.3) with a four-outcome categorical variable: no dissent (omitted), dissent by the median judge, dissent by the second-extremal judge, and dissent by the extremal judge:

$$\text{dissent}_p = \alpha \, \text{bipartisan}_p + \text{controls}_p + \gamma_{ct} + \varepsilon_{pct}, \tag{3.4}$$

[13] The location of the median judge influences the possible distances, e.g., a median judge located at 0 has a maximum distance of 1 from any other judge, while a median judge located at -0.5 has a distance of 1.5 to a judge located at 1.

$$\begin{aligned}\text{dissent}_p = \beta_1\,\text{extremal-dist}_p + \beta_2\,\text{2nd-extremal-dist}_p + \beta_3\,|\text{median}|_p \\ + \text{controls}_p + + \gamma_{ct} + \varepsilon_{pct}.\end{aligned} \tag{3.5}$$

Equation (3.4) is the analogue of Equation (3.2), and Equation (3.5) is the analogue of Equation (3.3). I estimate both new specifications as conditional multinomial logits—i.e., estimating the change in each type of dissent separately—and cluster robust standard errors at the circuit × year level. Estimates for these specifications are in Table C.3.[14] Columns (1)–(4) use JCS scores to determine the median, second-extremal, and extremal judge and as variables of interest on the right-hand side. Columns (5)–(8) use DIME-CF scores instead.

Contra the median voter prediction, and new to the literature on how judicial polarization affects outcomes, dissents from all three judges increase equally. The differences between the coefficients are not significant even at the 20% level in any of the full-sample specifications of columns (1), (2), (5), and (6). The magnitude of the differences is no more than a tenth the size of the overall increase in dissent rates by each judge. As with Table C.2, the effect is more clearly visible for JCS scores, where it is driven entirely by bipartisan panels. Using DIME-CF scores gives the same roughly equal increase in dissent rates, but the effect is weak enough on bipartisan panels to be marginally significant only at the 10% level.

These results offer a sharp contrast to existing models of judicial politics. Both the theoretical literature discussed in Section 3.2 and the model I specialize to my setting in Section 3.5.1 emphasize that median judges are much less likely to dissent, whether because they strategically join majority coalitions or because noise may place them further to the left or right than the data implies. Using the point estimates, dissents by the median judge increase by 1.7% less than dissents by the second-extremal judge and by 1.2% more than dissents by the extremal judge, for an average of 0.25% less than their more extreme colleagues overall. This value represents a dissent share of 33.2% compared to the proportional share of $33.\bar{3}$%. Re-computing this difference in 1,000 Bayesian bootstrap draws of the data gives a 95% confidence interval for the median dissent share ranging from 27.2% to 39.1%. Even at the lower bound of the confidence interval, on over 1,000 cases each year the two politically extreme judges form a coalition which excludes the median judge. That lower bound still only accounts for about one-fourth of the causal increase in dissents on bipartisan panels: when partisanship causes judges to fail to reach consensus, that failure is non-partisan in nature almost 75% of the time. To conclusively establish equal dissent rates, in the next section I provide two robustness checks using alternative estimation strategies, followed by an investigation of the effect's strength across time and case topics.

3.6.3 Equal Dissent Rates Are Robust

The models presented in Section 3.5 provide some theoretical foundations for believing that dissent rates may truly be equal across majority and minority partisans, as well as across median and extremal judges. However, one clear possible alternative is that ideology scores are merely a low-quality, noisy measure of partisan polarization. To rule out this explanation, recall that Table C.2 shows ideology scores are just as predictive of increased dissents as party, and the magnitude of that increase is practically significant. In addition, increased

[14] As with the earlier results on total dissents, these results are robust to considering Democratic- and Republican-majority bipartisan panels separately; see tables C.19–C.21 in Appendix C.5.1.

dissents are driven exclusively by the extremal judge's distance. If noise alone explained similarities in dissent rates, this distance should not affect the second-extremal judge's dissent rate, which should instead be driven by their own distance from the median. However, the second-extremal judge's distance has no greater effect on their own dissent rate than on the other judges, and that effect is generally not significantly different from zero.

To further alleviate the concern that noise drives the results, I provide two additional tests that dissents are indeed equal across majority-party/median judges and minority-party/extremal judges. The first is a binary-outcome specification at the judge level:

$$\text{Dissent}_{jp} = \alpha\, \text{judge-char}_{jp} + \text{controls}_p + \gamma_j + \gamma_{ct} + \varepsilon_{jpct}, \tag{3.6}$$

which I can estimate as a conditional logit with standard errors clustered at the judge level. In this specification, the additional subscript j indicates that I consider specifically dissents by, and characteristics of, judge j on panel p of circuit $\times$ year ct. That is, each observation of a panel becomes three observations of individual judges, and "judge-char" captures, e.g., whether or not judge j is a minority partisan on panel p. I use judge fixed effects γ_j to control for variation in individual judge characteristics. Results from this specification are in columns (1)–(3) of Table C.4.[15]

Before interpreting the results, I comment on a challenging issue of statistical power. To a first approximation, there are 200-500 cases observed on average per judge. Given that most judges serve for several decades, I can conservatively upper-bound each judge at 50 cases observed per circuit $\times$ year. Since dissent rates are only around 10%, this means a judge dissents fewer than five times in each random sample, too small to separate out noise in dissents from the effects of panel assignments. In contrast, the panel-level specification has an average of 300-500 cases per circuit $\times$ year and can thus detect more notable differences in the authorship of the corresponding ≈ 50 dissents.

With this caveat in mind, I can use the confidence interval to rule out any differences in dissents between minority- and majority-partisan judges larger than 0.1 percentage points, or about one-twentieth the size of the increase in dissents caused by switching a party-unanimous panel to a bipartisan one. The difference in dissents between median and extremal judges is similar, at most 0.15 percentage points. Clearly, the increase in dissents on bipartisan panels cannot be driven by markedly higher dissent rates by judges when they are in the partisan minority, nor is it driven by dissents from the ideological extremes.

Constructing a panel-level test of equal dissent rates by minority partisans is more challenging and requires carefully defining the counterfactual. In Table C.3, I take advantage of the fact that ideology scores are defined on both party-unanimous and bipartisan panels, i.e., extremal and median judges exist regardless of whether all three judges are Democrats or one of them is a Republican. Thus, I can ask how making the extremal judge more extreme affects the dissent rates of all three judges, using the full sample of panels where ideology scores exist in the data. I can also pose a well-defined question relating party to ideology scores: if one of the judges on a three-Democrat panel were made to be a Republican, would the extremal judge dissent more? However, a minority-party judge is (clearly) only well-defined on a bipartisan panel, so I cannot test whether a minority-party judge dissents more on bipartisan panels than party-unanimous ones.

[15] In Table C.22 (Appendix C.5.2), I use controls for party, age, and gender instead of judge fixed effects; results are qualitatively unchanged.

Since the dissent rate of majority-party judges is well-defined on both party-unanimous panels and bipartisan panels, I can estimate it and impute the dissent shares of majority- and minority-party judges. Let d be the dissent rate on bipartisan panels, which I can derive from the dissent rate u on party-unanimous panels and the logit coefficient estimated in Equation (3.2). I estimate a modified version of that specification where I replace the left-hand-side indicator for dissents with an indicator for majority-party dissents. Of course, on all party-unanimous panels the indicators will have the same value, so u will be unchanged. However, I obtain a new estimate d_{maj} of the dissent rate on bipartisan panels using this new indicator, reflecting dissents by majority partisans only. If $d_{\text{maj}}/d = 2/3$, then the dissent rate of majority partisans is exactly as would be expected if party has no effect on dissent rates; if instead $d_{\text{maj}}/d < 2/3$ (resp. $> 2/3$), then minority partisans are making a larger (resp. smaller) than proportional contribution to the change in dissent rates. I can perform a similar test to see whether the median judge (measured using either JCS or DIME-CF scores) makes up a greater-than-expected share of dissents on bipartisan panels compared to party-unanimous ones, but must compare the ratio d_{maj}/d to 1 instead of 2/3 because median judges exist on both party-unanimous and bipartisan panels. To form a confidence interval and perform statistical significance tests for these ratios, I compute them for 1,000 Bayesian bootstrap draws of the data; results are in columns (1)–(3) of Table C.5.

In all cases, the ratios are qualitatively close to the proportional values expected if party or score polarization did not affect dissent rates, though in five of the six cases I reject the null of exact equality. In particular, the ratio suggests that minority partisans make up a share of dissents about 0.4% larger than proportional, while median judges' share of dissents is about 0.4% less. This difference is minute; scaling by the approximately 10% of cases with dissents, it translates to dissenting 0.04 percentage points more often, essentially insignificant compared to the 2.0 p.p. increase in dissents on bipartisan panels.

Having established that, in the full sample, majority partisans and ideological-median judges dissent at essentially the same rate as minority partisans and ideological-extreme judges, I now confirm that this effect persists across various key subsets of the data. The share of dissents by each judge role (minority-party, median, second-extremal, or extremal) is approximately one-third in each year from 1970 to 2013, the full time period for which scores are available. This pattern is illustrated in figures C.18 and C.19 (Appendix C.5.2), which aggregate results across circuits for ease of presentation. The results across different legal topics are summarized in Table C.6, which presents coefficients equivalent to column (1) of Table C.3 estimated within each of the Bloomberg legal topics, i.e., the increase in dissent rates of the median, second-extremal, and extremal judge (as measured using JCS scores) when switching from party-unanimous to bipartisan panels. In particular, equality continues to hold for the three most common topics—criminal, due process, and economics and regulation—as well as for miscellaneous cases. The median judge dissents far less for civil rights cases, but far more for labor relations cases. Split-party panels have no more dissents than unanimous-party panels on First Amendment cases, but the sample is an order of magnitude smaller and the baseline dissent rate is 50% higher than for any other topic.

Returning to the comparison with judge-level specifications, the panel-level result clarifies an important subtlety about how a judge's individual distance from the median affects dissents. For panels where that judge is not extremal, their distance from the median has no effect on any judge's dissent probability. When plotting individual dissent probability as a

function of distance from the median, this result will manifest as a positive slope which is nearly zero for small distances from the median, where the individual judge is highly unlikely to be extremal, and increases with that distance—not because the effect is quadratic, but because the probability of being extremal and having any effect at all increases with distance. The effect on the overall probability of dissent is in fact three times larger than the effect on a judge's own probability of dissent. Both of these results are confirmed qualitatively in Figure C.2, which uses local polynomial regression to show dissent probability as a function of individual or extremal distance from median.

3.7 Judge Identity Influences Dissents

Prior literature has found that non-partisan demographic characteristics can influence outcomes and dissents [4, 50]. In this section, I explore whether my two main conclusions about partisanship—partisan polarization causes more dissents, but those dissents are not driven by the most-polarized outlier judge—hold for ideologically relevant non-partisan characteristics as well. For example, does taking a panel where all three judges share a law school (the equivalent of a party-unanimous panel) and replacing one member with a judge from a different law school (the equivalent of switching to a bipartisan panel) cause more dissents? If so, is the increase driven by a judge's increased likelihood of dissenting when they are an outlier, i.e., in the law school minority?

To answer this question, I replicate the four types of analysis from Section 3.6: the effect of mixed panels on total dissents (Table C.2 in Section 3.6, Table C.7 here), the effect of mixed panels on dissents by outlier judges (Table C.3 in Section 3.6, Table C.8 here), the effect of mixed panels on judge-level dissents (Table C.4), and the share of dissents by majority judges (Table C.5). Across three important non-partisan characteristics—law school, gender, and age—I find marginally statistically significant increases from the presence of outlier judges, but the magnitude of the point estimates is smaller than the increase caused by partisanship. I also provide suggestive evidence that dissents are somewhat concentrated among "outlier judges," e.g., the law school minority judge as described above. This combination of empirical facts is more consistent with both a classical median-voter-style framework where outliers directly cause dissents and the model of sincere dissents with noise I presented in Section 3.5.1.

3.7.1 Non-Partisan Characteristics Cause Some Dissents

I begin by following the approach of Equation (3.2) to test whether the presence of outlier judges (e.g., replacing a man on an all-men panel with a woman, or an oldest judge whose age is close to their colleagues with a much older judge) causally increases total dissents. In particular, I use specifications of the form

$$\text{dissent}_p = \alpha \, \text{panel-char.}_p + \text{controls}_p + \gamma_{ct} + \varepsilon_{pct}, \tag{3.7}$$

restricting the sample to contain only panels where the selected characteristic is relevant. As with Equation (3.2) in Section 3.6.1, I estimate each specification as a conditional logit and cluster robust standard errors at the circuit × year level. Results are in Table C.7.

Only gender causes a statistically significant change in dissent rates; however, all signs are as expected, suggesting that the presence of an outlier on a panel (whether that means mixed parties, law schools, or genders, or ages) increases dissents.[16] Effects are roughly one-third to one-fourth as large as the effect of partisanship. When estimating a single conditional logit model including indicators for all of these possible outlier effects (Table C.26 in Appendix C.5.3), the coefficient for party is essentially unchanged. All other coefficients, with the exception of those for gender outliers (e.g., replacing a man on an all-male panel with a woman), are slightly reduced in magnitude and become not statistically significant. These results suggest that outlier effects are present for non-partisan characteristics, but their effect on judges' tendency to disagree is small in comparison to the effect of partisanship.

3.7.2 Dissents are Somewhat Concentrated on Outlier Judges

The clearest test that the increase in dissents caused by partisan polarization is distributed equally across judges is Table C.3 (Section 3.6.2). That result separates the change in dissents on bipartisan panels according to a judge's role as median, second-extremal, and extremal using a partisan ideology score. To replicate this test for, e.g., law school, I need a continuous measure of law school similarity, paralleling how JCS and DIME-CF ideology scores measure the partisan similarity of two judges. Such a measure would allow me to define a "law-school-extremal" judge on each panel, analogous to the ideology-score-extremal judge.

As a first step in this direction, I use the semantic embedding vectors for each opinion to construct a simple measure of similarity. Each law school is represented by the average vector across the semantic embeddings for all opinions whose authors went to that law school, and the similarity between two law schools is simply the cosine similarity of their representing vectors. Because cosine similarity is symmetric, a panel therefore includes three similarity measures: the similarities of judges 1 and 2, 2 and 3, and 1 and 3. (If two judges share a law school, their similarity is 1.) The intuitive analogue of the ideology-score-extremal judge in this setting is therefore the judge who is *not* a member of the most similar pair.[17] Restricting to panels with either two or three unique law schools—since on panels with one unique law school all judges have the same similarity—I can thus estimate the specification,

$$\begin{aligned} \text{dissent}_p = &\text{has-three-schools}_p \\ &+ \text{controls}_p + \gamma_{ct} + \varepsilon_{pct}, \end{aligned} \tag{3.8}$$

where "dissent" on the left-hand side is a categorical variable capturing no dissent, dissent by the (law-school-) extremal judge, dissent by the second-extremal judge, and dissent by the median judge.[18] As with Table C.3, I estimate the specification as a conditional multinomial

[16] Particularly striking is the coefficient for panels with exactly two unique law schools; it flips sign from column (2), where these panels are more mixed than those with one unique law school, to column (3), where those panels are less mixed than those with three unique law schools.

[17] I can analogously define the second-extremal judge as the one who is not a member of the second-most similar pair. The remaining judge, who is a member of both the most and second-most similar pairs, is the median judge.

[18] Note that with two unique law schools, both the median and second-extremal judge are in the law school majorty; which is which is determined by their similarity to the law school minority judge.

logit and cluster standard errors at the circuit × year level. Results from this estimation are in column (1) of Table C.8.

Before discussing those results, I briefly describe the parallel estimation approach for my other two variables of interest, gender and age. To reproduce these results for gender, I could use (or re-construct) the measure of gender slant used in Ash et al. [4]; I could then test whether the increase in dissents on mixed-gender panels is concentrated among more gender-slanted judges. Though this exercise is straightforward, it requires additional data and extensive computations, so I leave it for future work. However, effects by age are easily computed using existing data: there are always youngest, median, and oldest judges on each panel, so I can test how the (not statistically significant) increase in dissents from the presence of an age outlier is distributed across judges. In particular, I estimate the specification

$$\begin{aligned} \text{dissent}_p = &\text{has-age-outlier}_p \\ &+ \text{controls}_p + \gamma_{ct} + \varepsilon_{pct}, \end{aligned} \tag{3.9}$$

where "dissent" on the left-hand side is a categorical variable capturing no dissent, dissent by the youngest judge, dissent by the age-median judge, and dissent by the oldest judge. As with the other parallel specifications, I estimate Equation (3.9) as a conditional multinomial logit and cluster standard errors at the circuit × year level. Results from this estimation are in columns (2)–(4) of Table C.8.

While the difference in coefficients for law school in column (1) of Table C.8 is not significant, the difference in point estimates is meaningful given the magnitude of the overall effect (columns (2) and (3) of Table C.7). In particular, the increase in dissents is 4.1% for the (law-school-) median judge compared to 2.0% and 2.3% for the second-extremal and extremal judges respectively. That is, dissents are reduced for outlier judges. However, this reversal is less surprising when considering that on a panel with two unique law schools, there is a sharply defined law school majority and minority; when that division is eliminated, the members of the majority are more affected by the removal of their likely coalition party than the already-isolated minority.

Effects for age are also mostly suggestive. The oldest judge sees a clear increase in dissents and the median judge appears to dissent less than the others, but the differences between the coefficients are again not statistically significant. Furthermore, the oldest judge actually dissents most often when there is a young outlier, rather than (as expected) when they are the old outlier themselves. Given the role of the senior judge in assigning opinion authorship, further analysis—and potentially more data—is needed to untangle the endogenous effects of opinion assignment and dissenting behavior.

I next return to the judge-level specification of Equation (3.6), which produced the result that switching an individual judge from the partisan majority to the partisan minority did not increase their dissent rate (column (1) of Table C.4). Here, I test the effect of switching a judge from the majority to the minority in a non-partisan characteristic—e.g., comparing the dissent rate of Judge A, a woman, on panels where she is the only woman to her dissent rate on panels where she is one of two women.[19] As a reminder, the relevant specification is

[19]For the effect of age, I separately test the effect of being the youngest, median, and oldest judge on the panel. Unlike ideology scores, where the direction of distance from the median is likely irrelevant, young and old judges may behave differently.

a conditional logit,

$$\text{dissent}_{jp} = \alpha \, \text{judge-char}_{jp} + \text{controls}_p + \gamma_j + \gamma_{ct} + \varepsilon_{jpct},$$

where "judge-char" is, e.g., an indicator for whether judge j is in the panel gender minority. To ensure that this comparison is not capturing a difference between, e.g., mixed-gender and gender-homogeneous panels, akin to the difference between bipartisan and party-unanimous panels, the sample in each column contains only mixed-characteristic panels for the appropriate characteristic.[20] Results are in columns (4)–(10) of Table C.4.

Noise is even more of an issue in this setting than in the partisan one, as shown by the dramatically smaller sample sizes; this issue is driven by the rarity of panels where all three judges share a law school (about 3.5% of the sample), as well as panels with two women (about 2.7% of the sample). Still, in both the version with controls for judge characteristics and the version with judge fixed effects, I can rule out large (5-10%) differences in dissent rates at the judge level for all non-partisan characteristics except for gender, where there is a 13.4% increase in dissents for women judges from being in the gender minority.

Lastly, aggregating back to the panel level, I repeat the comparison of panel-majority judges' behavior on unanimous and mixed panels, following the approach in Section 3.6.2. Specifically, I estimate the conditional logit specification

$$\text{dissent}_p = \alpha \, \text{mixed-char}_p + \text{controls}_p + \gamma_{ct} + \varepsilon_{pct}, \tag{3.10}$$

where the left-hand side is an indicator for the presence of a dissent by any judge on panel p and "mixed-char" is an indicator for, e.g., panel p being mixed-gender rather than gender-unanimous.[21] The estimated coefficient $\hat{\alpha}$ and baseline dissent rate on characteristic-unanimous panels u allow me to compute the dissent rate on mixed-characteristic panels, d. I then re-estimate the specification replacing the left-hand side with a new indicator "dissent_maj" which only tracks dissents by judges in the characteristic majority on panel p, and obtain a new dissent rate d_{maj}. If the ratio d_{maj}/d is less than (resp. greater than) the proportional level of 2/3—or 1, for characteristics present on both mixed and unanimous panels—then the characteristic majority contributes fewer (resp. more) dissents than is proportional. Results from this procedure for law school, gender, and age are in columns (4)–(10) of Table C.5. As in Section 3.6.2, standard errors and statistical significance are obtained by re-computing the ratio using 1,000 Bayesian bootstrap draws of the data.

Recall that majority partisans account for an almost-exactly-proportional 66.3% of the increase in dissents on bipartisan panels (column (1) of Table C.5. Here, for all characteristics except law school I am able to reject the null that dissent shares are proportional, i.e., as expected if that characteristic played no role in dissenting behavior. Indeed, for all characteristics except law school and women judges (where the sample size is extremely small, driven by the rarity of majority-women panels), both the difference from proportionality of the point estimate and the level of rejection of the null are larger than for any of the analogous partisan measures. In particular, gender-minority judges' share of dissents is 3.5%

[20] Recall that I replicate the results using judge-level controls in Table C.22 (Appendix C.5.3).

[21] For age, I use the indicator for a young outlier when looking at dissents by youngest judges, the indicator for an old outlier when looking at dissents by oldest judges, and the presence of any age outlier when looking at dissents by median judges.

larger than proportional; youngest judges' share is 1.7% smaller; median-age judges' share is 1.8% smaller; and oldest judges' share is 1.7% larger. Converting these ratios into total dissents, the gender-minority judge contributes 0.36 percentage points more dissents than the gender-majority judge, more than half of the 0.64 p.p. increase in dissents caused by mixed-gender panels. Oldest judges contribute about 0.15 p.p. more dissents, almost half of the 0.35 p.p. increase in dissents caused by having an old outlier. While these effects seem small in percentage terms, they represent hundreds of additional dissents per year and may affect judges' ability to reach consensus on highly impactful issues.

Overall, analyzing non-partisan judge characteristics in the same manner as partisanship shows two clear differences from the results presented in Section 3.6. First, the effect of non-partisan characteristics on total dissents is much more limited. Only the effect of gender is significantly different from zero at the 5% level, and while the signs of all other effects are as expected, the point estimates are only about one-fourth the size of the highly statistically significant effect of party. It therefore seems that partisanship is uniquely powerful in causing disagreements. The second difference is in which judges are affected by the increase in dissents. While that the effect of party is robustly distributed across not only majority and minority partisans but also extremal and median judges, the increase in disssents driven by mixed-characteristic panels is more concentrated among outlier judges. This effect is often weak, but given the reduced increase in total dissents, milder effects are anticipated. These results are consistent with classical models of judicial politics, e.g., the framework in Section 3.5.1. Even if the increase in total dissents and the localized effects on outlier judges were both approximately zero—which I cannot rule out in some cases—that pair of results would be consistent with a simple decision-making process which is not affected by these non-partisan traits. In contrast, the presence of a strong positive effect of partisanship on total dissents rules out the explanation that partisanship is simply irrelevant to appellate decisions.

3.7.3 Partisan and Non-Partisan Characteristics Have Limited Interactions

The results of the previous section make clear that the effect of partisanship on dissents is both stronger and more evenly distributed than that of other non-partisan judge characteristics. However, these effects may interact; a "double-minority" judge who is, e.g., a minority partisan and the only woman on a panel with two men, may behave differently than either a minority partisan in the panel gender majority or a majority partisan who is in the panel gender minority. In Table C.25 (Appendix C.5.3), I interact the outlier effects of Table C.7 with whether the outlier judge is a majority or minority partisan. Double-minority law school judges do seem to dissent more than their single-minority outlier counterparts, but double-minority age and gender judges do not.

To more directly analyze these relationships, I follow the approach of Equation (3.2) and consider the "cross-effect" of increased dissents from panel partisanship on characteristic-

majority and characteristic-minority judges.[22] I estimate a conditional multinomial logit,

$$\text{dissent}_p = \alpha \, \text{bipartisan}_p + \text{controls}_p + \gamma_{ct} + \varepsilon_{pct}, \tag{3.11}$$

where the left-hand side is a categorical variable with outcomes no dissent, dissent by the characteristic-minority judge, and dissent by the characteristic-majority judge. (As with the estimation of Equation (3.2), I include controls for panel-level variables and cluster robust standard errors at the circuit × year level.) Specifically, I compare the law school minority and majority on panels with two unique law schools, the gender minority and majority (as well as separately considering a gender-minority man and gender-minority woman) on mixed-gender panels, and all three possible age roles (youngest, age-median, and oldest). Results are in Table C.9.

Beginning with law school, the difference in increased dissents between minority and majority judges is not statistically significant, but the qualitative results are sharply different from those of Table C.3, where the point estimates across median and extremal judges were essentially identical. Here, I find a 10.9 percentage point (p.p.) larger increase in the dissent rate for the law school minority judge (36.5% compared to 25.6%), which represents 0.9 p.p. more cases with dissents. Of course, this effect only applies on about 25% of bipartisan panels (those with two unique law schools), and thus could account for about one-eighth of the total increase in dissents caused by partisanship. That is, if bipartisan panels merely activated latent disagreements across law schools, causing differential increases in dissents by law school minority and law school majority judges—with no other effects—bipartisan panels would have about 0.25 p.p. more dissents than party-unanimous ones (one-eighth of the result from Section 3.6.1).

Turning next to gender, the difference between majority and minority judges is not statistically significant when combining men and women judges, but is statistically significant when separately examining the comparison between majority and minority men and majority and minority women. The increase in dissent rate is 125.7 p.p. larger for men in the gender minority (151.6% compared to 25.9% in the gender majority), representing 11.2 p.p. more cases with dissents. The increase in dissent rate is actually 32.3% smaller for women in the gender minority (13.5% compared to 45.7% in the gender majority), representing 3.1 p.p. more cases with dissents. That is, women are more affected by partisan polarization when in the gender majority than when in the gender minority. Since panels with exactly two women make up only about one-eighth of mixed-gender panels, the pooled effect of being in the gender minority is slightly larger than the pooled effect of being in the gender majority. Benchmarking these effects against the 2.0 p.p. more cases with dissents on bipartisan panels, panels with men in the gender minority are about 2% of bipartisan panels and could thus account for about one-tenth of the total increase in dissents caused by partisanship; panels with women in the gender minority are about 17% of bipartisan panels and could thus account for one-quarter of the total increase in dissents caused by partisanship.

Finally, examining age, the difference between youngest, age-median, and oldest judges is again not statistically significant; however, the point-estimated increase in dissent rate for

[22] Given the lack of power in the judge-level specification when looking at each characteristic individually, with this data it is infeasible to estimate, e.g., the dissent rate of double-minority party and law school judges compared to single-minority judges along each dimension.

youngest judges is 4.0 p.p. larger than for median judges (36.6% compared to 32.6%); the difference between median and oldest judges is minimal (32.6% compared to 33.0%). The former represents 2.7 p.p. more cases with dissents, and the latter only 0.4 p.p. more cases with dissents. Since the sample for this specification includes all bipartisan panels, the effect of age could in fact cover the whole effect of partisanship. However, caution is needed in interpreting this result because of the senior judge's role in assigning opinion authorship. This assignment may be endogenous, e.g., youngest judges know that a senior judge of the opposite party is unlikely to give them authorship, and thus dissent more on bipartisan panels where such a judge can be present. Still, it supports a framework where non-partisan disagreements make up a larger share of total dissents than partisan ones.

As seen in this section, any study of the interaction between partisanship and other judge characteristics is limited by the dramatically reduced number of "doubly-mixed" panels. As such, these results merely suggest that partisanship activates other latent disagreements rather than showing it conclusively. However, I can provide stronger evidence that these non-partisan characteristics have quantitatively important effects on opinion content—and thus justify disagreement by outlier judges—through a closer analysis of opinion text. This complementary approach, which I lay out in the next section also allows me to work with the full sample of opinions, increasing statistical power and more robustly establishing the influence of law school and gender in particular.

3.8 Using Natural Language Processing To Understand Opinion Text

Dissents are rare, but judges may have reason to be dissatisfied with an opinion even if they are willing to sign it—or may even sign an opinion they disagree with for strategic reasons (as in the model extension of Appendix C.3.2). Because these disagreements do not produce differences in observed outcomes, they cannot be measured using standard metadata. However, by analyzing opinion text using natural language processing (NLP) tools, I can quantify differences in opinions that suggest potential areas of disagreement. This approach completes the picture of whether partisanship is directly motivating disagreements or simply contributing to an environment where judges who differ along non-partisan dimensions are less likely to reach consensus.

I focus on one important way in which an outlier judge might affect a panel's discussion: bringing new legal topics to the table, or contributing to a far-ranging discussion which covers more topics in total. Focusing on broad topics abstracts away from other dimensions along which judges might disagree—such as the presence of emotive language, or the valence with which particular arguments are discussed—and may not fully capture the richness of legal decision-making. However, these topics are still an important feature of decisions, determining which areas of law they are most likely to influence through precedent and citations. They are also an input into the decision to dissent: most dissents include an additional topic not present in the lead opinion, suggesting that the dissenting judge felt a particular issue was not given sufficient consideration. Thus, by looking at the distribution of topics across panel types and individual characteristics, I can obtain a picture of when an opinion is likely to be

unsatisfying to a given judge without having to rely on the presence of an actual dissent.

Using state-of-the-art NLP transformer models to identify and measure the prevalence of legal topics in the full sample of opinion text, I find evidence that a panel's partisan composition has little effect on the topics discussed in lead opinions.[23] Author party has some influence, but is lesser in both range and magnitude to the effect of gender, while author age is largely irrelevant. The effect of law school on opinion content is more challenging to measure and requires a less supervised approach using local similarity measures, but appears to be comparable in magnitude to the effect of gender.

To obtain these results, I follow a standard fine-tuning pipeline using a transformer model pre-trained on legal documents (LegalBERT from Chalkidis et al. 23) and predict the legal topics hand-coded for each case by Bloomberg Law.[24] In Section 3.8.1, I follow in spirit Gentzkow et al. [39], who use predicted author party as a measure of polarization in Congressional speech, and compare predicted topic distributions across various panel compositions and author characteristics. I supplement analysis of the distribution of individual topics with another important dimension of potential heterogeneity: the breadth of topics covered, which I measure as the number of "uniform major topics," with probability larger than 1/8 (i.e., larger than if drawn uniformly at random) and the number of "Bayesian major topics" with higher predicted probability than their prior probability.[25] In Section 3.8.2, I move beyond predicted topics and use novel measures based directly on semantic embeddings to study whether author characteristics affect the overall similarity of opinions.

3.8.1 Analyzing Predicted Legal Topics

A partisan-minority judge is less likely to be chosen as lead opinion author under both random author selection and author selection by similarity to the senior judge, whose identity is itself quasi-random. Thus, if there is a large difference in opinion topics across parties, a partisan-minority judge will often disagree with the content of the lead opinion. The sharpest comparison across parties is of all-Democratic to all-Republican panels; these panels are more likely to be ideologically homogeneous and ensure there can be no moderating effect of author assignment (e.g., a Republican senior judge writes an opinion even though the panel is majority Democratic). On the other hand, the comparison of Democratic-majority bipartisan panels to Republican-majority bipartisan panels is especially crucial for a minority partisan's decision to dissent. In the analysis that follows, I present both comparisons, as well as a comparison of the transition from party-unanimous to bipartisan panels.

Focusing first on party-unanimous panels, I use the following specification:

$$\begin{aligned}\mathbb{P}(topic)_p = \beta_1 \,&\text{panel-type}_p \\ &+ \text{controls}_p + \gamma_{ct} + \varepsilon_{pct},\end{aligned} \tag{3.12}$$

[23] In this section, I focus on the time period 1970–2013, when Democratic and Republican positions on these topics—civil rights in particular—are more consistent, and data is fully available for all judge characteristics of interest.

[24] For details of the prediction task and performance metrics, see Appendix C.4.

[25] Tables C.28 and C.29 in Appendix C.6.1 show that the latter two computations are robust to scaling those thresholds. Figures C.20–C.22 show the full distributions of both kinds of major topics across panels with different numbers of Democrats, women, and unique law schools.

where "panel-type" captures whether the panel is all-Democratic or all-Republican. I estimate a conditional logit model for each individual topic probability (since they are bounded between 0 and 1) and a conditional Poisson model for uniform and Bayesian major topics (since they are integer counts), and cluster robust standard errors at the circuit × year level. Because author assignment may be endogenous, I focus on panel partisanship, which is randomly assigned. However, to allow qualitative comparisons of party to gender and age, in Appendix C.6.2 I estimate this specification with controls for author age and gender (Table C.30). Results from the estimation of Equation (3.12) are in the first row of Table C.10.

Recall that criminal, due process, and economics and regulation make up about 85% of the Bloomberg-coded ground truth topic labels. On the first two of those, there are no significant differences across parties. However, Democratic panels are less likely to discuss economics and regulation. Among the remaining topics, Democratic panels are more likely to discuss both First Amendment and miscellaneous issues. They are also about 15% more likely to include an additional major (Bayesian) topic. This result fits well with Democrats' increased tendency to discuss the broad miscellaneous topic.

These effects represent changes in the average share of text devoted to each topic, rather than shifts from, e.g., an opinion entirely about economics and regulation to one about civil rights. However, interpreting them in the latter way allows me to benchmark them against the overall share of dissents. Democratic panels cover the First Amendment 0.17 percentage points (p.p.) more often, economics and regulation 1.13 p.p. less often, and miscellaneous topics 0.88 p.p. more often. These estimates are generally smaller than, but are comparable in magnitude to, the 2.0 p.p. increase in dissents on bipartisan panels compared to party-unanimous ones.

I next break down the effect of a more Democratic bipartisan panel using three related comparisons. First, I look only at Democratic-majority panels, and compare unanimous ones to bipartisan ones; I then do the same for Republican-majority panels, and finally compare topics across all bipartisan panels separated by majority party. For each comparison, I follow the specification in Equation (3.12) and again estimate a conditional logit model for all individual topic probabilities and a conditional Poisson model for uniform and Bayesian major topics, clustering robust standard errors at the circuit × year level.[26]

The effects of partisanship on Democratic- and Republican-dominated panels qualitatively match those in Table C.30, with the additional Democratic judge having a stronger effect on majority-Democratic panels than majority-Republican panels (where they have essentially no effect). Strikingly, but perhaps not surprisingly given the lack of influence from partisanship on observable disagreement, the effect of switching from a Republican-majority bipartisan panel to a Democratic-majority one is essentially null. The only statistically significant coefficient is a marginal increase in discussion of First Amendment issues (0.06 p.p.), and the other point estimates are generally smaller than in any of the other comparisons.

More speculatively, I can use the results in tables C.30–C.33 (Appendix C.6.2) to compare the influence of partisanship and other judge characteristics. I caveat this comparison by noting that if author assignment is endogenous to these characteristics, then the results

[26]Tables C.31–C.32 in Appendix C.6.2 replicate the analysis with controls for author party, age, and gender. Table C.34 in the same appendix replicates the analysis with author fixed effects, essentially estimating the difference in, e.g., opinions written by Republican Judge A when on a panel with two Democrats to opinions written by Judge A when on a panel with one more Republican and one Democrat.

may be biased; however, most of the literature on author assignment suggests that it is driven primarily by idiosyncratic features like judge workloads and domain-specific knowledge. Across the four comparisons, the effect of gender is almost universally larger topic-by-topic than the effect of party, while the effect of age is largely null. In particular, women judges tend to discuss civil rights and miscellaneous topics more frequently, but labor relations and economics and regulation topics less frequently. The magnitudes of these effects are generally between 0.5 p.p. and 2 p.p., as substantial as the increase in dissents on bipartisan panels. Also of interest in this setting with controls is the ability to compare panel effects (through the coefficient on panel characteristics) to author effects (through the coefficient on author party). In Tables C.31–C.33, the effect of author partisanship is generally stronger than, but in the same direction as, the effect of panel partisanship. However, the effect of author partisanship is almost always smaller than the effect of author gender.

To establish conclusively that authorship is more important than panel peer effects, or that gender is definitively more important than party, further research is needed to establish when author assignment is quasi-random. However, these results strongly support the idea that party has little effect on topic choice. Additionally, the effect of gender compared to party is especially clear on bipartisan panels (Table C.33 in Appendix C.6.2), where control of the majority has no effect on any individual topics or the number of major topics, while the effect of author party is dominated by that of gender in all categories but one (where it is essentially equal). While these results do not speak directly to the *existence* of dissents, they suggest that a woman judge in the gender minority on a panel is more likely to disagree with the content of the lead opinion than a panel-minority partisan, providing evidence for the influence of non-partisan characteristics on the *motivation* to dissent. A missing variable in this analysis is law school training, which is challenging to compare under this approach because it takes many unordered values. In the next section, I address this gap by directly using the learned text embeddings rather than the transformer model's already-powerful reduced-form predictions.

3.8.2 Analyzing Opinion Similarity

A major benefit of modern transformer approaches to NLP is that these models encode a complete embedding representation of any text input as a high-dimensional vector, which is then used for classification and prediction tasks as in the previous section. Since the model I use is fine-tuned to represent opinions in topic space, closeness of these embeddings reflects similarity in legal topics, but at a much finer-grained level than the predictions can convey. To gain sharper insight into which judge characteristics have the most influence on legal topics in opinion text, I select the 1,000 nearest neighbors for each opinion—about 0.4% of the full sample—by cosine similarity (the standard measure of closeness in embedding space) and assess the properties of these neighbors. This measure provides a more nonparametric approach to the motivating question of the previous section: should minority partisans be worried that an opinion written by a majority partisan will be different from the opinion they would have liked to write?

Using author party as an example, I describe below the steps for constructing my measure of by-characteristic similarity:

1. For each opinion, compute the share of neighboring opinions for which the author's party matches the neighbor's author's party.

2. For each party in the sample, compute the overall share of opinions authored by judges of that party.

3. Normalize the share of matches by subtracting the appropriate mean and dividing by the entropy.[27] E.g., for a Republican-authored opinion, subtract the share of Republican-authored opinions in the sample from the share of neighboring Republican-authored opinions and divide by the entropy of authorship party.

4. Compute the distribution of normalized scores and the average normalized score across all opinions.

The normalized match scores allow me to compare similarity along dimensions like party, where there are few categories and the sample is balanced, to dimensions like gender, where there are still few categories, but the sample is highly unbalanced, and dimensions like law school, where there are many categories and the sample is unbalanced. I present the results, averaging across opinions in Table C.11; full distributions of normalized and de-meaned match scores are in figures C.23–C.28 (Appendix C.6.3).[28]

The score for legal topic serves as a useful benchmark; not only was this variable the target for fine-tuning the model, it also displays the clearest clustering patterns when viewing a low-dimensional representation of the embedding vectors. Figure C.3 compares this by-topic visualization to one which colors opinions by the number of Democrats on the panel; the difference in visible clusters is immediately apparent. While these visualizations necessarily sacrifice some of the information contained in the full embedding vectors, they provide a first suggestion that partisanship will not markedly alter the pattern of nearby opinions.

The normalized match score for legal topic is an order of magnitude larger than the scores for law school or gender, both of which are roughly of the same magnitude as the match scores for author party and number of Democrats, but opposite in sign. For the three partisan characteristics, the match scores are negative, indicating that along the chosen dimension opinions are less similar to actual neighbors than if those neighbors were randomly selected. For the three non-partisan characteristics, the match scores are positive, indicating higher similarity than under a random choice of neighbors. While not directly comparable due to the differing numbers of outcomes for each variable, the de-meaned coefficients have the most straightforward interpretation: they represent the percentage of similar neighbors in excess of the population average. For example, in column (5) of Table C.11 the average opinion is 2.7% more likely to have a neighbor authored by a judge from the same law school than if that neighbor had been drawn by random chance. Focusing on this column in particular, the most-represented law school in the sample makes up only 15.7% of the observations, so this effect represents at least a 20% increase over the base rate, and much

[27] I use entropy rather than variance as in a standard z-score because several of the variables, e.g., law school, are categorical rather than numeric.

[28] Results on average match scores are robust to choosing the 100 or 10,000 nearest neighbors (tables C.35 and C.36, Appendix C.6.4) instead of the 1,000 nearest neighbors, as well as to considering only neighbors with different authors (tables C.37 – C.39, Appendix C.6.4).

more for less-frequent law schools. Also striking are the share of opinions with de-meaned match scores greater than 0, i.e., the share of opinions whose neighbors are more similar to them than is the sample as a whole. Slightly less than half of opinions written by Democratic judges have more Democratic-authored neighbors than Republican-authored neighbors. In contrast, almost 70% of opinions authored by a judge from a given law school or of a given gender have more shared-characteristic neighbors than random chance would predict. All results hold across different numbers of neighbors—including in the 10,000-neighbor sample, where the neighborhood is about 4% of the full sample—and regardless of whether or not those neighbors include other opinions by the same author.

The results on legal topic clearly show that local similarity in topic space is well-captured by semantic embeddings, further validating the NLP approach to this question. Using those embeddings to compare partisan and non-partisan judge characteristics shows that opinions' neighbors are essentially random with respect to party, corroborating the reduced-form results on topic distributions in the previous section. Importantly, this new approach allows me to quantify the degree to which shared law schools affect local similarty. While the effect is an order of magnitude smaller than that of legal topic, it is still comparable in magnitude to the effect of gender. Given the strong reduced-form findings for the effect of gender on opinion text, these results suggest that law school training is an important driver of how opinions are written. Thus, judges in the law school minority are likely to find themselves disagreeing with the lead opinion even when they are unwilling to dissent.

3.9 Conclusion

Appellate court rulings are in many ways an ideal setting for research in political economy: they are high-volume, high-impact, and richly documented, but present important challenges due to the technical nature of many appellate rulings. In this work, I complement existing work on partisan outcomes of appellate panels by documenting and explaining a key empirical pattern: partisanship causes an increase in dissents, but that increase is distributed evenly across all judges on a panel, not just the partisan minority. This combination of facts adds a new dimension to the well-documented claims of increased judicial polarization in the 21st century. I also show that it is unique to partisanship; other important judge characteristics have a more limited effect on total dissents and concentrate that effect among outlier judges. While the behavior along those non-partisan dimensions can be rationalized by classical median-voter or sincere dissent frameworks with sufficient noise in preferences, I show that even with very large levels of noise the effect of partisanship cannot be explained by either of those two approaches.

I instead describe a framework where multidimensional judges have preferences over majority coalitions based on the similarity of the judges in the coalition. To explain the results, this framework requires non-partisan effects to dominate partisan ones. I provide suggestive evidence that this dominance does indeed occur by showing that dissents on bipartisan panels are higher among law school and gender outliers, though reduced sample size limits the strength of these findings. Moving beyond observed dissents, I contribute novel applications of NLP tools to capture the legal topics discussed in appellate opinions, and show that the effect of gender is indeed larger than that of party. The flexibility and

informativeness of text embeddings also allow me to compare the full content of opinions, not just reduced-form measures like predicted topics or partisanship. Local similarity metrics in topic space show that judges from the same law school cover similar topics, and benchmarks the magnitude of that effect favorably against those of other demographic variables. Supported by both classical metadata and novel text data, I conclude that non-partisan characteristics do indeed have a strong influence on opinions, and thus on judges' decision to dissent.

This work focused on appellate courts largely in isolation; a key direction for future work is to consider them as part of the hierarchical U.S. court system and combine appellate court data with information about federal trial courts and the Supreme Court. In line with recent work by Cohen and Dehejia [30], I find that bipartisan panels are more likely to uphold a decision by a 2-1 vote than to reverse it by the same margin, and that unanimous reversals are slightly decreased.[29] That work finds that appellate judges are more deferential to lower court judges of their own party; combined with the intuitive assumption that reversing a decision requires greater agreement and a stronger argument than upholding one, I obtain my empirical result. An interesting test of this reasoning would be to combine the appellate court data with information about the party (and possibly ideology score) of the lower-court judges to test whether the dissenting judge in a 2-1 reversal is more likely to be of the same party as the lower court judge. I could also use text similarity measures combined with automatic NLP-based sentiment analysis to assess the degree to which an appellate ruling favorably discusses and extends the lower court's reasoning. In general, the rapidly advancing state of the art in NLP offers a variety of avenues along which to extend my results. One immediately promising methodological approach is to directly augment text with metadata. This can be done either by concatenating a BERT-like transformer-generated embedding with encoded information about panel and judge characteristics and jointly optimizing the resulting vector for classification tasks, or by including those characteristics in few-shot prompts to a large language model. Beyond these richer embeddings, approaches like sentiment analysis and automated argument detection can provide more explainable descriptions of how opinions differ across judge characteristics and bring insights into the intricacy of their reasoning. These innovations, and surely many others, represent not only important contributions to research in political economy, but also key steps in advancing public understanding and monitoring of a well-functioning judiciary.

[29] See Table C.27 in Appendix C.5.3 for details.

Appendix A

Additional Material for "Persuasion with Ambiguous Receiver Preferences"

A.1 Omitted Proofs for Section 1.4

A.1.1 Maxmin Persuasion and MPC Games

I first show that any Nash equilibrium strategy for Sender in the MPC game described in Section 1.4 solves the maxmin persuasion problem of Equation (1.1):

Lemma 1. *Consider the MPC game with tie-breaking against Sender, where Sender's choice set is defined as*

$$\mathcal{G} = \{\textit{cdf } G \textit{ over } [0,1] \,|\, G \textit{ is an MPC of } F\},$$

and Nature's choice set is defined as

$$\mathcal{T} = \{\textit{cdf } T \textit{ over } [0,1] \,|\, T \textit{ is an MPC of } (1-r^*)\,\delta_0 + r^*\delta_1\}.$$

$G^* \in \mathcal{G}$ is a Nash equilibrium strategy for Sender if and only if G^* solves the maxmin persuasion problem of Equation (1.1).

Proof. I first show that Sender's utility from the maxmin persuasion problem is the same as from an analogous minmax problem. By Proposition 1 of Kleiner et al. [52], both $\mathcal{G}$ and $\mathcal{T}$ are compact and convex. Because this result uses the norm topology, both spaces are metric spaces, hence Hausdorff spaces.

The functional of Equation (1.1) is linear in both distributions, so it is convex in G and concave in T. For fixed $G \in \mathcal{G}$, it is also lower semicontinuous on T: the result follows from lower semicontinuity of the indicator function (which applies for $q > r$, any $q \in [0, r]$ produces the same value) and application of Fatou's Lemma.

Therefore, I can apply Theorem 2 of Fan [36] to state that

$$\sup_{G\in\mathcal{G}} \left\{ \min_{T\in\mathcal{T}} \int\int \mathbf{1}(q > r)\, dG(q)\, dT(r) \right\} = \min_{T\in\mathcal{T}} \left\{ \sup_{G\in\mathcal{G}} \int\int \mathbf{1}(q > r)\, dG(q)\, dT(r) \right\}.$$

Because $\mathcal{G}$ is compact, I can in fact replace the supremum on the left-hand side of the equation with a maximum, giving precisely the expression in Equation (1.1).[1] It is then clear that Sender's utility with simulatenous moves in the MPC game must be equal to her utility in the maxmin persuasion problem.

I now prove the "if" portion of the lemma. Let G^* solve the maxmin persuasion problem. Since the maxmin and minmax utilities for Sender are equal, the game has a value, and both Sender and Receiver have strategies that guarantee them at least the value. G^* is by definition such a strategy; let T^* be such a strategy for Receiver. It must be that the pair (G^*, T^*) guarantees each player exactly the value of the game because it is zero-sum: if either player's utility were strictly above the value, then the other's would be strictly below it. Thus G^* is a best response to T^*, since no other strategy gives Sender strictly higher utility (or T^* would not guarantee Receiver the value). Similarly, T^* is a best response to G^*. Thus G^* is a Nash equilibrium in the MPC game for Sender.

For the "only if" portion, let G^* be a Nash equilibrium strategy in the MPC game for Sender. Then the Nash equilibrium payoff for Sender results from taking Nature's best response to G^*. Since Nature's payoff is the opposite of Sender's, that payoff is therefore Sender's minimum utility from G^*. Thus a Nash equilibrium distribution for Sender has the same payoff in the MPC game and the maxmin persuasion problem, and that utility is precisely equal to Sender's maximum utility in the maxmin persuasion problem, so G^* solves the maxmin persuasion problem by definition. □

This equivalence does not rely on the support of the prior F in the maxmin persuasion problem, and thus suggests that results from other maxmin persuasion models may be applied to solve richer MPC games.

I next show that different tie-breaking assumptions in the persuasion context can be re-interpreted as restrictions on the domain of distributions each player can choose in the MPC game context.

Lemma 2. *Consider a two-player MPC game G where the support of Sender's and Receiver's chosen distributions must lie weakly above a common lower bound and weakly below a common upper bound, and tie-breaking selects Receiver as the winner if there is a tie. This game is equivalent to a game F which is identical to G except for the following two changes:*

1. *The support of Receiver's distribution is not bounded above—Receiver may choose any mean-preserving contraction that obeys the lower bound.*

2. *Tie-breaking is even—in the case of a tie, a winner is randomly chosen.*

Proof. The game F is (modulo simultaneous moves, which Lemma 1 shows are irrelevant) the same as the maxmin persuasion problem. There, Sender persuades Receiver type $r < 1$ by generating a posterior $q_r^0 = r$; with unfavorable tie-breaking, she must generate $q_r^\varepsilon = r + \varepsilon$ for arbitrary $\varepsilon > 0$. Since posteriors must lie in $[0, 1]$, posteriors q_1^ε are infeasible and Sender can never persuade Receiver type $r = 1$. As $\varepsilon \to 0$, the effect on the Bayes-plausibility constraint from replacing any q_r^0 with q_r^ε vanishes, allowing a Sender facing unfavorable tie-breaking to

[1] I cannot replace the supremum on the right-hand side with a maximum, and indeed the results for MPC games are often stated using limits of sequences of distributions.

match her utility with favorable tie-breaking (and thus for any intermediate tie-breaking rule) for interior Receiver types, but not for type $r = 1$. Thus Sender's utility is affected by the tie-breaking rule if and only if she chooses a posterior distribution with an atom at $q = 1$.

The MPC game G, where the chosen distributions must have support in $[0, 1]$ for Sender and in $[0, \infty)$ for Receiver, means that even under favorable tie-breaking for Sender, Nature can generate Receiver types $r_q^\varepsilon = q + \varepsilon$ and keep Sender's utility to the same level as with Receiver-favoring tie-breaking. In particular, Nature can generate type $r_1^\varepsilon = 1 + \varepsilon$, which it would not be able to do if constrained by the upper bound. Thus, a Nash equilibrium of G is equivalent to one of F. □

Since tie-breaking against Sender allows me to work with a well-defined minimizing Receiver type distribution for each posterior distribution, I choose this rule. Thus, combining both lemmas, I may apply existence and uniqueness results from MPC games with arbitrary tie-breaking rules and choice set

$$\mathcal{G} = \{\text{cdf } G \text{ over } [0, 1] \,|\, G \text{ is an MPC of } F\},$$

for Sender and

$$\mathcal{T}' = \{\text{cdf } T \text{ over } \mathbb{R}_+ \,|\, T \text{ is an MPC of } (1 - r^*)\,\delta_0 + r^*\delta_1\}.$$

for Nature.

A.1.2 Characterizing Sender's Optimal Distribution

Given the equivalence result of the prior section, a Nash equilibrium of the MPC game in Hart [43], which the author calls a "Captain Lotto game," provides a solution to the maxmin persuasion problem of Equation (1.1). Thus the strategy of Player B in Theorem 4 of that work now gives an optimal posterior distribution for Sender. Theorems 4 and 5 of Amir [1] show that the Nash equilibrium strategy for Player B in the Captain Lotto game is unique when $\pi \leq 1/2$, and therefore so is Sender's optimal posterior distribution. To complete the proof of Proposition 1, I replace the sufficient condition for Nash equilibrium when $\pi > 1/2$ in Theorem 10 of Amir [1] with a necessary and sufficient condition for optimality of Sender's chosen posterior distribution.

I begin by showing that Sender's utility can be expressed as a function of $\bar{G}$, the concavification of G:

Lemma 3. *Consider the maxmin persuasion problem of Equation (1.1) and let $\bar{G} : [0, 1] \to [0, 1]$ be the concavification of G, i.e., the infimum over the set of concave functions $H : [0, 1] \to [0, 1]$ satisfying*

$$H(q) \geq G(q) \;\; \forall \;\; q \in [0, 1].$$

Then the following equality holds:

$$\min_{T \in \mathcal{T}} \int \int \mathbf{1}(q > r)\, dG(q)\, dT(r) = 1 - \bar{G}(r^*).$$

Proof. Manipulating the bounds of integration to rewrite Sender's objective function from Equation (1.1) gives

$$\int_{[0,1]} \left(\int_{[0,1]} \mathbf{1}(q > r)\, dG(q) \right) dT(r) = \int_{[0,1]} \left(\int_{[r,1]} 1\, dG(q) \right) dT(r)$$
$$= \int_{[0,1]} (1 - G(r))\, dT(r).$$

Then the minimzation portion of the problem can be written as

$$\max_{T \in \Delta([0,1])} \int G(r)\, dT(r) \quad \text{s.t.} \quad \int r\, dT(r) = r^*,$$

where I have dropped the constant, rewritten the min as a max, and explicitly included the mean restriction to highlight the similarity to a Bayesian persuasion problem. In this case, the Receiver type r fills the role of "posterior belief," Nature's utility from a realized Receiver type is $G(r)$, and the "prior" is the distribution with support $\{0, 1\}$ and mean r^*. This final point follows from the observation in Section 1.3 that when the prior distribution has binary support, the Bayes-plausibility constraint is the same as a mean restriction. Thus by Corollary 2 of Kamenica and Gentzkow [48], Nature's utility is given by $\bar{G}$. the concavification of G over the interval $[0, 1]$, evaluated at the prior mean r^*. Flipping the sign again, Sender's utility is $1 - \bar{G}(r^*)$. □

My necessary and sufficient condition is an immediate consequence of this result:

Lemma 4. *Let $\pi > 1/2$. Then a posterior distribution G^* is optimal for Sender if and only if $\mathbb{E}_{G^*}[\omega] = \pi$ and $G^*(q) \leq q \ \forall \ q \in [0, 1]$. More than one distribution satisfying this condition always exists.*

Proof. As established in Section 1.3, the only constraint on a feasible distribution G for Sender is that $\mathbb{E}_G[\omega] = \pi$; I show that the second constraint is both necessary and sufficient for optimality.

Assume $G^*(q) \leq q \ \forall \ q \in [0, 1]$. Then the function $U(q) = q$ upper-bounds G^* and is concave. Since U is the pointwise-smallest concave function on $[0, 1]$ passing through the point $(1, 1)$, it must therefore be the concavification of G^*, and Sender's utility from G^* is $1 - U(r^*) = 1 - r^*$. Because Nature may always choose a Receiver type distribution T with $\text{supp}(T) = \{0, 1\}$, Sender's utility from any posterior distribution is no more than $1 - r^*$ (the probability that Receiver type $r = 0$ is drawn from T). Thus G^* attains the upper bound and is optimal for Sender. There are at least two such distributions for any $\pi > 1/2$. The first is given by solving $\pi = n/(n + 1)$ for n and setting $G^*(q) = q^n$. The second is given by

$$G^*(q) = \begin{cases} 0 & q \in [0, 2\pi - 1), \\ (q + 1 - 2\pi)/(2 - 2\pi) & q \in (2\pi - 1, 1]. \end{cases}$$

Therefore an optimal distribution always exists and is non-unique.

Now assume G^* is optimal for Sender; then, since I have just shown an optimal distribution exists, it must be that $1 - \bar{G}^*(r^*) = 1 - r^*$. But the only weakly positive concave function

H on $[0,1]$ satisfying $H(1) = 1$ and $H(r^*) = r^*$ is $U(q) = q$. Any distinct concave function must have slope greater than 1 at r^*—any less and it would fail to pass through the point $(1,1)$—and must therefore have $H(0) < 0$. Therefore $\bar{G}^* = U$ and $G^*(q) \leq q \ \forall \ q \in [0,1]$. □

This lemma completes the proof of Proposition 1. However, the theorems I reference rely on lengthy computations of Sender's utility under different strategy profiles. In the next section, I provide a clearer geometric proof that does not rely on MPC games and instead highlights the usefulness of concavification.

A.1.3 An Alternative Proof of Proposition 1

The concavification result of Lemma 3 means that Sender's utility from any posterior distribution G is a convex function of r^*. It can therefore be lower-bounded by a line tangent to that function through the fixed r^* in the maxmin persuasion problem. The key step of my alternative proof of Proposition 1 is to show that if G gives Sender a higher utility than the optimal distribution G^*, then that tangent line implicitly defines a cdf whose mean is greater than π. Because the tangent line lies below the function $1 - G$, it must therefore be that G itself has a mean greater than π, and thus G is not a Bayes-plausible posterior distribution.

Towards establishing this result, consider *upper-truncated uniform posterior distributions* (henceforth UTUs), a class of posterior distributions which place mass $x \geq 0$ on posterior $q = 0$, equal mass on all posteriors $q \in (0, r_h]$ for some $r_h \leq 1$, and no mass on posteriors $q \in (r_h, 1]$. I can use Bayes-plausibility to solve for the unique value of r_h corresponding to a given x, so that a UTU is fully characterized by x:

$$\pi = \int q \, dG_x(q) = \int 1 - G_x(q) \, dq = \frac{(1-x)\, r_h(x)}{2}$$
$$\Leftrightarrow r_h(x) = \frac{2\pi}{1-x},$$

Since r_h is uniquely determined by x, I denote a UTU by G_x. The following lemma shows that a single choice of x is optimal among all UTUs and can be written as a closed-form function of r^*:

Lemma 5. *Let $\pi \leq 1/2$. Then if $r^* \leq \pi$, Sender's unique optimal UTU is G_0; if $\pi \leq r^* \leq 1/2$, it is $G_{1-\pi/r^*}$; and if $1/2 \leq r^*$ it is $G_{1-2\pi}$.*

Proof. By construction, any UTU G_x is concave and is therefore equal to its concavification $\bar{G}_x$. By Lemma 3, the utility from a UTU G_x is therefore

$$1 - \bar{G}_x(r^*) = 1 - G_x(r^*) = \left\{ (1-x) - r^* \frac{(1-x)^2}{2\pi} \right\}_+ .$$

The first-order condition in x for the expression in brackets is

$$-1 + r^* \frac{1-x}{\pi} = 0 \Leftrightarrow x_{\text{FOC}} = 1 - \frac{\pi}{r^*}.$$

The bracketed expression is increasing in x when $x < x_{\text{FOC}}$ and decreasing in x when $x > x_{\text{FOC}}$. Since $x \in [0, 1-2\pi]$, if $r^* < \pi$ the constrained optimal solution is $x^* = 0$ and if $r^* > 1/2$ the constrained optimal solution is $x^* = 1 - 2\pi$; otherwise the optimum is the interior solution $x^* = x_{\text{FOC}} = 1 - \pi/r^*$. □

I now prove two lemmas describing the relationship between the UTU $G_{1-2\pi}$ and the function $1-\bar{G}$ derived from an arbitrary posterior distribution G. The first establishes that if, for some posterior distribution G, the function $1-G$ falls below $1-G_{1-2\pi}$ at some mean Receiver type q, Sender's utility from G remains below her utility from $G_{1-2\pi}$ for all higher Receiver types:

Lemma 6. *Let G be a cdf on $[0,1]$. Then if there is $q \in [0,1)$ such that*

$$1-G(q) < 1-G_{1-2\pi}(q),$$

then it is also the case that

$$1-\bar{G}(q') < 1-G_{1-2\pi}(q') \ \ \forall \ \ q' \in [q,1).$$

Proof. The proof is by contradiction. Assume there is q such that

$$1-\bar{G}(q) \leq 1-G(q) < 1-G_{1-2\pi}(q),$$

but that there is $q' \in [q,1)$ such that

$$1-\bar{G}(q') \geq 1-G_{1-2\pi}(q').$$

Since $1-\bar{G}(q) < 1-G_{1-2\pi}(q)$ but $1-\bar{G}(q') \geq 1-G_{1-2\pi}(q')$, it must be that there is $q_1 \in [q,q']$ where the slope of $1-\bar{G}$ is strictly greater than that of $1-G_{1-2\pi}$. But because G and $G_{1-2\pi}$ are cdfs and $1-\bar{G}$ is weakly positive,

$$1-\bar{G}(1) = 0 = 1-G(1) = 1-G_{1-2\pi}(1),$$

so there must be $q_2 \in [q',1]$ where the slope of $1-\bar{G}$ is weakly less than that of $1-G_{1-2\pi}$. Then $q_1 \leq q_2$ but the slope of $1-\bar{G}$ at q_1 is strictly greater than at q_2, violating convexity of $1-\bar{G}$, and thus concavity of $\bar{G}$. □

The next lemma describes features of $1-\bar{G}$ when the posterior distribution G weakly improves on Sender's utility from $G_{1-2\pi}$:

Lemma 7. *If $G \neq G_{1-2\pi}$ is a cdf such that*

$$1-\bar{G}(r^*) \geq 1-G_{1-2\pi}(r^*) \text{ and } \int q\, dG(q) = \pi,$$

then the slope[2] of $1-\bar{G}$ at r^* is strictly less than the slope of $1-G_{1-2\pi}$ at r^*.

Proof. I first show that there is $q_d \in (r^*,1]$ such that

$$1-\bar{G}(q_d) \leq 1-G(q_d) < 1-G_{1-2\pi}(q_d).$$

[2]Because $1-\bar{G}$ is convex, it is continuous on $(0,1)$ and its left and right derivatives are always well-defined. The function $1-G$ for any UTU G is also continuous with well-defined left and right derivatives. When referring to the slope or to a tangent line I consider the right derivative.

Note that for G to be distinct from $G_{1-2\pi}$, there must be some posterior $q_d \in [0,1]$ where $1-G(q_d) \neq 1-G_{1-2\pi}(q_d)$. It cannot be the case that

$$1-G(q) \geq 1-G_{1-2\pi}(q) \ \ \forall \ \ q \in [0,1] \ \ \text{and} \ 1-G(q_d) > 1-G_{1-2\pi}(q_d).$$

If that is the case, then because G is a cdf, it is right-continuous, and therefore fixing $\varepsilon > 0$ there is $\delta(\varepsilon) > 0$ such that

$$1-G(q') > 1-G(q_d) - \varepsilon \ \ \forall \ \ q' \in [q_d, q_d + \delta(\varepsilon)).$$

Since the slope of $1-G_{1-2\pi}$ is no greater than 0, setting $\varepsilon \in (0, G_{1-2\pi}(q_d) - G(q_d))$ ensures that

$$1-G(q') > 1-G_{1-2\pi}(q_d) \geq 1-G_{1-2\pi}(q') \ \ \forall \ \ q' \in [q_d, q_d + \delta(\varepsilon)).$$

Therefore there is a non-degenerate interval where $1-G > 1-G_{1-2\pi}$, and by assumption $1-G \geq 1-G_{1-2\pi}$ everywhere on $[0,1]$, so integrating the inequality gives a violation of Bayes-plausibility:

$$\int q \, dG(q) = \int 1-G(q) \, dq > \int 1-G_{1-2\pi}(q) \, dq = \int q \, dG_{1-2\pi}(q) = \pi.$$

Thus by contradiction there must be $q_d \in [0,1]$ such that

$$1-\bar{G}(q_d) \leq 1-G(q_d) < 1-G_{1-2\pi}(q_d).$$

By Lemma 6, since $1-\bar{G}(r^*) \geq 1-G_{1-2\pi}(r^*)$, there is no $q \in [0, r^*)$ where $1-G(q) < 1-G_{1-2\pi}(q)$. Thus it must be that

$$1-G(q) \geq 1-G_{1-2\pi}(q) \ \ \forall \ \ q \in [0, r^*],$$

and therefore $q_d \in (r^*, 1]$.

The claim now follows by the argument in Lemma 6. Since $1-\bar{G}(r^*) \geq 1-G_{1-2\pi}(r^*)$ and $1-\bar{G}(q_d) < 1-G_{1-2\pi}(q_d)$, there is $q' \in [r^*, q_d]$ where the slope of $1-\bar{G}$ is strictly less than that of $1-G_{1-2\pi}$. But since $\bar{G}$ is concave, $1-\bar{G}$ is convex and its slope cannot increase as q decreases; the slope of $1-\bar{H}$ at r^* must therefore be strictly less than that of $1-G_{1-2\pi}$ at r^*. □

The implication is vacuous for $r^* \leq 1/2$, where there are no posterior distributions that meet the conditions; however, even in that case the result is central to a proof by contradiction.

With these three lemmas in hand, I now provide an alternative proof of the case $\pi \leq 1/2$ in Proposition 1:

Lemma 8. *If $\pi \leq 1/2$, Sender's unique optimal posterior distribution is as follows:*

- *If $r^* \leq \pi \leq 1/2$,*
$$G^*(q) = U[0, 2\pi].$$
- *If $\pi \leq r^* \leq 1/2$,*
$$G^*(q) = \left(1 - \frac{\pi}{r^*}\right) \delta_0 + \frac{\pi}{r^*} U[0, 2r^*].$$

- *If* $\pi \leq 1/2 \leq r^*$,

$$G^*(q) = (1 - 2\pi)\,\delta_0 + 2\pi\, U[0,1].$$

Proof. The proof is by contradiction. Let G be a proposed alternative posterior distribution that delivers weakly greater utility for Sender than G^*. By Lemma 3 (to define the utility from each posterior distribution) and Lemma 5 (since G^* is a UTU, it must be uniquely optimal among UTUs), it is the case that

$$1 - \bar{G}(r^*) \geq \bar{G}^*(r^*) = 1 - G^*(r^*) \geq 1 - G_{1-2\pi}(r^*).$$

Consider the line L that is tangent to $1 - \bar{G}$ at r^*.[3] Because $\bar{G}$ is convex and weakly positive (recall that the line $\ell(q) = 0$ is convex and lower-bounds $1 - G$), it is lower-bounded by $L_+(q) = \max\{L(q), 0\}$. Furthermore, by Lemma 7, the slope of L is less than that of $1 - G_{1-2\pi}$, so it must be that

$$1 \geq 1 - G(0) \geq 1 - \bar{G}(0) \geq L_+(0) > 1 - G_{1-2\pi}(0) = 2\pi.$$

For any $x \in [0, 1-2\pi]$, there is a corresponding UTU G_x with $G_x(0) = x$. Since $1 - L_+(0) \in [0, 1 - 2\pi]$, there exists an UTU—call it G_{alt} for alternative—with $1 - G_{\text{alt}}(0) = L_+(0)$. If $G_{\text{alt}} \neq G^*$, then because G^* is uniquely optimal among UTUs, it must be that

$$L_+(r^*) = 1 - \bar{H}(r^*) \geq 1 - \bar{G}^*(r^*) > \bar{G}_{\text{alt}}(r^*) = 1 - G_{\text{alt}}(r^*).$$

Then, because L and $1 - G_{\text{alt}}$ intersect at $q = 0$ but L is greater than $1 - G_{\text{alt}}$ at $q = r^*$, it must be that the slope of L is strictly greater than the slope of the strictly downward-sloping portion of $1 - G_{\text{alt}}$; therefore in fact

$$L_+(q) \geq 1 - G_{\text{alt}}(q) \;\; \forall \;\; q \in [0,1] \quad \text{and} \quad L_+(q') > 1 - G_{\text{alt}}(q') \;\; \forall \;\; q' \in (0, r^*].$$

Integrating the expression and using the fact that L_+ lower-bounds $1 - \bar{G}$, which in turn lower-bounds $1 - G$, it is the case that

$$\begin{aligned}\int q\, dG(q) = \int 1 - G(q)\, dq \geq \int 1 - \bar{G}(q) dq \geq \int L_+(q)\, dq \\ > \int 1 - G_{\text{alt}}(q)\, dq = \int q\, dG_{\text{alt}}(q) = \pi.\end{aligned}$$

The first and penultimate equalities are both from integration by parts, and the final equality is because all UTUs (including G_{alt}) are Bayes-plausible by construction. Therefore G violates Bayes-plausibility and is not a valid alternative distribution.

Even when $G_{\text{alt}} = G^*$, it is still the case that, whenever

$$L_+(r^*) = 1 - \bar{G}(r^*) > 1 - \bar{G}^*(r^*) = 1 - G^*(r^*),$$

the slope of L is greater than the slope of the strictly downward-sloping portion of $1 - G^*$. In this case, $L_+(q) > 1 - G^*(q) \;\; \forall q \;\; \in (0, r^*]$ and

$$\begin{aligned}\int q\, dG(q) = \int 1 - G(q)\, dq \geq \int 1 - \bar{G}(q) dq \geq \int L_+(q)\, dq \\ > \int 1 - G^*(q)\, dq = \int q\, dG^*(q) = \pi,\end{aligned}$$

[3] Recall that if r^* is a kink point of $1 - \bar{G}$, I use the right derivative of $1 - \bar{G}$ to define the slope.

just as before. Thus G again violates Bayes-plausibility.

If instead $G_{\text{alt}} = G^*$ but now $L_+(r^*) = 1 - G(r^*)$, it must be the case that L and the strictly downward-sloping portion of $1 - G$ have the same slope, so in fact

$$L_+(q) = 1 - G^*(q) \ \ \forall \ \ q \in [0, 1].$$

Then then there are two possible cases. The first is trivial:

$$1 - G(q) = L_+(q) = 1 - G^*(q) \ \ \forall \ \ q \in [0, 1],$$

so that G is not a deviation at all. In the second, there must be some $q \in [0, 1]$ so that $1 - G(q) > L_+(q)$; recall that L_+ lower-bounds $1 - G$, and thus the direction of the inequality is known. Because G is a cdf, it is right-continuous, and therefore fixing $\varepsilon > 0$ there is $\delta(\varepsilon) > 0$ such that

$$1 - G(q') > 1 - G(q) - \varepsilon \ \ \forall \ \ q' \in [q, q + \delta(\varepsilon)).$$

Since the slope of L_+ is no greater than 0, setting $\varepsilon \in (0, 1 - G(q) - L_+(q))$ ensures that

$$1 - G(q') > L_+(q) \geq L_+(q') \ \ \forall \ \ q' \in [q, q + \delta(\varepsilon)).$$

Therefore there is a non-degenerate interval where $1 - G > L_+$, and $1 - G \geq L_+$ everywhere on $[0, 1]$, so integrating the inequality gives

$$\int q \, dG(q) = \int 1 - G(q) \, dq > \int L_+(q) \, dq = \int 1 - G^* dq = \int q \, dG^*(q) = \pi,$$

as desired. Having covered both the case $G_{\text{alt}} \neq G^*$ and the case $G_{\text{alt}} = G^*$, I have shown that in all cases H violates Bayes-plausibility and therefore, by contradiction, G^* is uniquely optimal. □

The full proof of Proposition 1 without reference to MPC games is therefore obtained by combining Lemmas 4 and 8.

A.1.4 Results with Alternative Tie-Breaking

The difference in Sender's optimal posterior distribution from the standard Bayesian persuasion problem is clearly driven by Nature's ability to tailor a worst-case Receiver type distribution to Sender's particular disclosure strategy, but may also be affected by the ability to generate a Receiver type who is unconvinced even when the state is surely $\omega = 1$. By following Lemma 2, I can obtain Sender's optimal posterior distribution under favorable tie-breaking, and show exactly when the choice of tie-breaking rule is influential:

Corollary 1 (Adapted from Theorem 4 of Hart 43)**.** *Let supp$(F) = \{0, 1\}$ and let ties be broken in favor of Sender.*
One optimal posterior distribution for Sender's is as follows:

- *If $r^* \leq \pi \leq 1/2$,*

$$G^*(q) = U[0, 2\pi].$$

- *If* $\pi \leq r^* \leq 1/2$,
$$G^*(q) = \left(1 - \frac{\pi}{r^*}\right)\delta_0 + \frac{\pi}{r^*}\,U[0, 2r^*].$$
- *If* $1/2 < r^*$,
$$G^*(q) = (1-\pi)\,\delta_0 + \pi\,\delta_1.$$
- *If* $r^* \leq 1/2 < \pi$,
$$G^*(q) = (2-2\pi)\,U[0,1] + (2\pi - 1)\,\delta_1.$$

Proof. I use the MPC game representation of the maxmin persuasion problem. The result follows directly from Player A's equilibrium strategy in Theorem 4 of Hart [43], where I replace the use of ε-approximating distributions, which are not needed in my setting, with the exact upper bound of 1 on posterior beliefs. Since the alternate tie-breaking rule is not the focus of this work, I do not provide a full characterization of other optimal posterior distributions; as Amir [1] shows, attempting a full characterization through the connection to MPC games becomes complex. □

In this case, Sender sometimes takes advantage of favorable tie-breaking and places an atom at posterior $q = 1$, exploiting Nature's inability to generate a skeptical Receiver type for that posterior belief. This choice allows Sender to obtain a utility higher than $1 - r^*$, since even if the Receiver type is $r = 1$, they are now persuaded whenever posterior $q = 1$ is realized. However, creating this atom tightens the Bayes-plausibility constraint, so if neither Sender nor Nature's constraint is slack enough to allow frequent realizations of 1, Sender uses the same approach as with unfavorable tie-breaking. Thus when the probability of the high state and the mean Receiver type are both small, Sender's maxmin utility remains strictly below her utility with even the most unfavorable prior belief about Receiver types, regardless of whether tie-breaking is favorable or not.

In the maxmin persuasion context, it seems natural to break ties either entirely in favor of or entirely against Sender. Those rules allow me to interpret a Receiver of type r either as the highest Receiver type who is convinced by posterior belief $q = r$ or the lowest Receiver type who is not convinced by that belief, respectively. However, if the MPC game is interpreted as competitive persuasion, as in Boleslavsky and Cotton [14], then it also seems reasonable to consider breaking ties evenly, so as to favor neither player.[4] This choice is equivalent to requiring that both distributions have support in $[0, 1]$; in that case the optimal posterior distribution (derived without uniqueness in Hart 43 and with uniqueness in Boleslavsky and Cotton 14 and Amir 1) is as follows:

Corollary 2 (Lemma 3 of Boleslavsky and Cotton 14)**.** *Let* $supp(F) = \{0, 1\}$ *and let ties be broken evenly.*
The unique optimal posterior distribution G^* *for Sender's is as follows:*

[4]When the players are persuading a Receiver about a common state of the world, as in Au and Kawai [6], it is reasonable to also require $\pi = r^*$. In Boleslavsky and Cotton [14], the players are schools convincing a Receiver about the binary ability of a student drawn from a school-specific distribution, so $\pi \neq r^*$ represents one school producing more high-type students on average.

- *If* $r^* \leq \pi \leq 1/2$,
$$G^*(q) = U[0, 2\pi].$$
- *If* $\pi \leq r^* \leq 1/2$,
$$G^*(q) = \left(1 - \frac{\pi}{r^*}\right) \delta_0 + \frac{\pi}{r^*} U[0, 2r^*].$$
- *If* $1/2 \leq \pi$ *and* $r^* \leq \pi$,
$$G^*(q) = \frac{1-\pi}{\pi} U[0, 2-2\pi] + \frac{2\pi - 1}{\pi} \delta_1.$$
- *If* $1/2 \leq r^*$ *and* $\pi \leq r^*$,
$$G^*(q) = \left(1 - \frac{\pi}{r^*}\right) \delta_0 + \frac{\pi}{r^*} \left(\frac{1-r^*}{r^*} U[0, 2-2r^*] + \frac{2r^* - 1}{r^*} \delta_1\right).$$

Proof. This result appears verbatim in Boleslavsky and Cotton [14], with Sender as Player A when $r^* \leq \pi$ and Player B when $r^* \geq \pi$. □

Finally, note for general interest that in the MPC game when both players' feasible distributions have domain $\mathbb{R}_+$ (and are mean-preserving contractions of binary support distributions), the unique solution is the same as the cases $r^* \leq \pi \leq 1/2$ and $\pi \leq r^* \leq 1/2$ of Corollary 2, with the relationship between π and r^* determining which case applies. The solution when $\pi = r^*$, so that the constraints are symmetric, first appears in Bell and Cover [10], and also appears in Myerson [64]. The solution for the asymmetric case first appears in Sahuguet and Persico [68], and also appears in Hart [42].

A.2 Omitted Proofs for Section 1.5

A.2.1 Properties of DTUs

To begin, I describe DTUs in more detail. The uniform portion of the DTU (between the lower and upper truncations) has slope β, which I refer to as the slope of the DTU. The line $L(q) = \beta q + y$, which forms that uniform portion, intersects the vertical axis at y; I refer to this value as the intercept of the DTU. To derive a relationship between β, y, and ℓ, I use the fact that Bayes-plausibility requires $\mathbb{E}_G[\omega] = \pi$. This condition immediately imposes the restriction that $\ell \in [0, \pi]$; using simple geometry to compute the integral of a DTU's cdf and set it equal to $1 - \pi$ shows that

$$\beta(\ell, y) = \frac{(\pi - y\ell) - \sqrt{(\pi - y\ell)^2 - \ell^2 (1-y)^2}}{\ell^2}.$$

This expression is continuously differentiable for $\ell \in (0, \pi]$ and $y \in [0, 1)$. Fixing ℓ, $\beta(\ell, y)$ is injective and decreasing in y. Fixing y, $\beta(\ell, y)$ is injective and increasing in ℓ, attaining a maximum of $\beta(\pi, y) = (1-y)/\pi$. While $\beta(0, y)$ is not defined using the expression above, the limit from the right exists:

$$\lim_{\ell \to 0^+} \beta(\ell, y) = \lim_{\ell \to 0^+} \frac{(1-y)^2}{(\pi - y\ell) + \sqrt{(\pi - y\ell)^2 - \ell^2 (1-y)^2}} = \frac{(1-y)^2}{2\pi}.$$

I thus define $\beta(0, y) = (1-y)^2/(2\pi)$ explicitly. For $y \in [0, 1 - 2\pi]$, $\beta(0, y)$ is the slope of the UTU with intercept y. When $y > 1 - 2\pi$, there is no corresponding UTU; instead, the lower bound of interest is $\beta(\ell, y) = 1 - y$, the slope that satisfies $G(1) = 1$.[5] The assumption $y > 1 - 2\pi$ implies $1 - y \in ((1-y)^2/(2\pi), (1-y)/\pi)$, so the lower bound is attained at an interior $\ell \in (0, \pi)$; I call this value $\ell_y^{\min}$. Because the function $\beta(\ell, y) - (1-y)$ is continuously differentiable, the Implicit Function Theorem ensures that I can write $\ell_y^{\min}$ as a continuously differentiable function of y.

The concavification of a DTU is easy to compute: so long as the slope of the line through $(0, 0)$ and $(\ell, \beta(\ell)\ell + y)$ is weakly less than $\beta(\ell, y)$, the concavification will be

$$\bar{G}_y^\ell(q) = \begin{cases} (\beta(\ell, y)\, \ell + y)/\ell & q \in [0, \ell), \\ G_y^\ell(q) & q \in [\ell, 1]. \end{cases}$$

That condition is simply

$$\frac{\beta(\ell, y)\, \ell + y}{\ell} = \beta(\ell, y) + \frac{y}{\ell} \geq \beta(\ell, y),$$

which always holds since $y \geq 0$ and $\ell \geq 0$. Thus the concavification of a DTU is composed of two upward-sloping line segments with a kink at ℓ and a constant line segment in the region of the upper truncation.

[5] This is the desired lower bound because any cdf H over $[0, 1]$ must satisfy $H = 1$, and I wish to use DTUs to upper-bound other feasible probability distributions.

A.2.2 Intercept-Optimal DTUs

Given a value of the mean Receiver type r^* and a fixed intercept y, I show the existence of a well-defined and unique DTU that provides Sender's highest utility among all DTUs with an intercept of y. Since y is fixed, for this section I drop the dependence on y from all functions.

Lemma 9. *Given $r^* \in (0,1)$ and $y \in [0,1)$, there is a well-defined DTU $G_y^{\beta(y)}$ with lower truncation length ℓ_y^* that maximizes Sender's utility among all Bayes-plausible DTUs with intercept y.*

Proof. Let $V_y \subset [0,\pi]$ be the set of ℓ such that a DTU with lower truncation ℓ and intercept y is Bayes-plausible. I first show that V_y is closed; since it is clearly also bounded, V_y is therefore compact. To do so, I define the function

$$v(x,\ell) = \int_0^x F(q)\,dq - \int_0^x G_y^\ell(q)\,dq$$

for some DTU G_y^ℓ with intercept y and lower truncation ℓ. This function captures the value of the Bayes-plausibility integral constraint for G_y^ℓ at $x \in [0,1]$. Clearly $v(0,\ell) = 0$, and $v(1,\ell) = 0$ because $\mathbb{E}_F[\omega] = \mathbb{E}_{G_y^\ell}[\omega] = \pi$.

At any x, the integral of G_y^ℓ on $[0,x]$ is continuous in ℓ. This result is obvious for $x \neq \ell$ (since $G_\ell(q)$ is continuous in ℓ at those points) and holds for $x = \ell$ because the left and right limits as $x \to \ell$ are both 0. Therefore $v(x,\ell)$ is also continuous in ℓ for fixed x, since it depends on ℓ only through that integral. If G_y^ℓ is not Bayes-plausible, then (since it satisfies $\mathbb{E}_{G_y^\ell}[\omega] = \pi$ by construction) there must be some $x_{\text{neg}} \in (0,1)$ for which $v(x_{\text{neg}},\ell) < 0$. Because $v(x_{\text{neg}},\ell)$ is continuous in ℓ, there is $\varepsilon > 0$ such that for any ℓ' in a ε-neighborhood of ℓ, $v(x_{\text{neg}},\ell') < 0$. Therefore any $G_y^{\ell'}$ is not Bayes-plausible, so $U \subset [0,\pi]$, the set of ℓ where Bayes-plausibility fails, is open. Since $V_y = [0,\pi] \setminus U$, it must be that V is closed.

By Lemma 3, Sender's utility from a DTU is given by

$$u_S(r^*,\ell) = 1 - \begin{cases} (\beta(\ell) + y/\ell)\, r^* + y & \ell < r^*, \\ G_y^\ell(r^*) & \ell \geq r^*. \end{cases}$$

This function is continuous in ℓ on $[0,\pi]$. Since $\beta(\ell)$ is continuous in ℓ on $[0,\pi]$, each of the two piecewise portions of u_S are clearly continuous in ℓ; it remains only to check the case $\ell = r^*$. But because the left and right limits as $\ell \to r^*$ exist (by continuity of each piecewise portion) and are equal (by construction of u_S), u_S is continuous at $\ell = r^*$ as well. Therefore the image of V under u_S must be compact, and thus contains a well-defined maximum, which is attained by some (possibly multiple) $\ell \in V$. □

Unlike in the binary-state setting, it is not possible to solve analytically for $G_y^{\beta(y)}$. However, appropriate sufficient conditions can ensure that $G_y^{\beta(y)}$ is both unique and slope-minimizing among Bayes-plausible DTUs with intercept y:

Lemma 10. *Fix $y \in [0,1)$ and $r^* \in (0,1)$. There is a unique and well-defined DTU G_y^{sm} that has minimal slope among all Bayes-plausible DTUs with intercept y. If $y = 0$ or $r^* \in [\pi,1)$, then the y-optimal DTU $G_y^{\beta(y)}$ equals G_y^{sm}*

Proof. Fix $y \in (0,1)$. By Lemma 9, the set V_y of values of ℓ such that G_y^ℓ is Bayes-plausible is closed, and the function $\beta(\ell, y)$ is continuous and monotonic in ℓ for fixed y, so there is a unique $\ell_{sm} \in V_y$ such that $\beta(\ell_{sm}, y) = \inf_{\ell \in V_y} \beta(\ell, y)$.

Now I show that either of the conditions provided in the lemma are sufficient for the slope-minimizing DTU to be optimal. First fix $y = 0$. Then $\beta(\ell) + y/\ell = \beta(\ell)$, so $u_S(r^*, \ell) = 1 - G(r^*)$; that is, there is no kink at ℓ in Sender's utility from DTUs with intercept 0. Thus Sender's utility from G_y^ℓ is strictly greater than her utility from $G_y^{\ell'}$ if and only if $\beta(\ell) < \beta(\ell')$. By Lemma 9, there exists a DTU $G_0^{\beta(0)}$ with lower truncation length ℓ_0^* that maximizes Sender's utility among all Bayes-plausible DTUs with intercept 0. No other Bayes-plausible DTU can have a strictly smaller slope, since then it would deliver a strictly higher utility. But no other Bayes-plausible DTU can have the same slope, $\beta(\ell_0^*)$, since there can be no $\ell' \neq \ell_0^*$ where $\beta(\ell) = \beta(\ell_0^*)$. Therefore all other Bayes-plausible DTUs have strictly larger slope, and so $G_0^{\beta(0)}$ satisfies both (1) and (2).

If instead $r^* \in [\pi, 1)$, then similarly $u_S(r^*, \ell) = 1 - G(r^*)$; since $\ell \in [0, \pi]$, r^* surely lies weakly above ℓ. The argument is then the same; a DTU is utility-maximizing if and only if it is slope-minimizing, Lemma 9 guarantees the existence of a utility-maximizing DTU, and the injectivity of the map from ℓ to $\beta(\ell)$ guarantees uniqueness. □

A.2.3 Simplifying the Integral Constraint

Let U_{r^*} be the set of utilities attained by any y-optimal DTU:

$$U_{r^*} = \left\{ u_S(r^*, \ell, y) \,|\, G_\ell^y = G_y^{\beta(y)} \text{ for some } y \in [0,1) \right\},$$

where I restore the dependence on y in u_S, since y is no longer fixed. That set is a subset of $[0,1]$, and is therefore bounded, so $\sup U_{r^*}$, Sender's supremum utility over all y-optimal DTUs (and thus over all DTUs) is well-defined and contained in the closure of U_{r^*}. Further restrictions on F and r^* provide sufficient conditions for U_{r^*} to be closed, and thus for the maximum to exist. In order to state these sufficient conditions, I first prove Lemma 11. In this proof, I again drop the dependence on y from all functions since y is fixed, but note important changes in the argument for different values of y.

Lemma 11. *Let G_y^{sm} be the DTU with the minimal slope among all Bayes-plausibile DTUs with intercept y, and let ℓ_y^{sm} be its lower truncation length. If $y \in [0, 1-2\pi]$, then the minimal interior q where $G_y^\beta(q) = F(q)$, call it $q_1(\ell, y)$, is well-defined and ℓ_y^{sm} satisfies*

$$\int_0^{q_1(\ell_y^{sm}, y)} F(q)\, dq = \int_0^{q_1(\ell_y^{sm}, y)} G_y^{sm}(q)\, dq$$

and

$$\int_0^x F(q)\, dq > \int_0^x G_y^{sm}(q)\, dq \qquad \text{if and only if} \qquad x \in (0, q_1(\ell_y^{sm}, y)) \cup (q_1(\ell_y^{sm}, y), 1).$$

If instead $y \in (1-2\pi, 1)$, then either the two conditions above hold or ℓ_y^{sm} equals the minimum lower truncation length ℓ_y^{min}.

Proof. By Lemma 10, there exists a unique minimal-slope Bayes-plausible DTU with intercept y.

Because of the shape of F, the equation $L(q) = \beta(\ell)\, q + y = F(q)$ has at most two solutions with $q \in (0, 1]$. In particular, if the slope of L is such that it lies completely above F in $(0, 1]$, then there are no solutions in that interval; if the slope of L is such that it is tangent to F, then there is one;[6] and if the slope of L is less than that of the tangent to F through y, there are two.

Consider a DTU $G_y^{\beta(\ell)}$ with lower truncation length ℓ. If $L(q) \geq F(q)\ \forall\ q \in (0, 1]$—that is, L is either tangent to F at a point q_t or lies entirely above F—then this DTU satisfies Bayes-plausibility. The function $v(x, \ell)$, which gives the value of the Bayes-plausibility integral constraint for $G_y^{\beta(\ell)}$ at some $x \in [0, 1]$, is weakly decreasing whenever $L(x) \geq F(x)$.[7] Thus $v(x, \ell)$ is weakly decreasing for all $x \in (\ell, 1)$. Since $v(1) = 0$, it must therefore be that $v(x, \ell) \geq 0\ \forall\ x \in (\ell, 1)$; of course $v(x, \ell) \geq 0\ \forall\ x \in [0, \ell]$, so in fact $v(x, \ell) \geq 0$ everywhere in $[0, 1]$ and Bayes-plausibility is satisfied.

The case where L intersects F twice in $(0, 1]$ will form the bulk of the proof. In particular, let q_1 be the smallest $q \in (0, 1]$ such that $\beta(\ell)\, q + y = F(q)$, and let q_2 be the largest.[8] By the Implicit Function Theorem, since the function $\beta(\ell)\, q + y - F(q)$ is continuously differentiable in all variables, I can write q_1 and q_2 as continuous functions of ℓ. Note that because of this definition, q_1 and q_2 are both well-defined (and satisfy $q_1 = q_2$) if $\beta(\ell)\, q + y$ is tangent to F, as well as for all smaller values of ℓ. I now address two-intersection DTUs by focusing on the cases $q_1(\ell) > \ell$ and $q_1(\ell) \leq \ell$.

If $q_1(\ell) > \ell$, then $G_y^{\beta(\ell)}(q) < F(q)$ for $q \in (0, \ell) \cup (q_1, q_2)$, but $G_y^{\ell}(q) > F(q)$ for $q \in [\ell, q_1) \cup (q_2, 1)$ (there is equality at $q \in \{0, q_1, q_2, 1\}$). Therefore if

$$v(q_1(\ell), \ell) = \int_0^{q_1(\ell)} F(q)\, dq - \int_0^{q_1(\ell)} G_y^{\ell}(q)\, dq \geq 0 \tag{A.1}$$

then $v(q) \geq 0\ \forall\ q \in [0, 1]$ and Bayes-plausubility is satisfied. Given the increasing and decreasing behavior of $v(x, \ell)$, it is clear that

$$v(q_1(\ell), \ell) = \min_{q \in (0,1)} v(q).$$

Therefore if a DTU violates Bayes-plausibility, it must be because $v(x, \ell) < 0$ for some $x \in (0, 1)$, which in turn implies that $v(q_1(\ell), \ell) < 0$. Thus when $q_1(\ell) > \ell$, Equation (A.1) is a necessary and sufficient condition for a DTU to be Bayes-plausible. Furthermore, if the inequality is strict for some ℓ, then because $v(q_1(\ell), \ell)$ is continuous in ℓ, it is also strict for $\ell - \varepsilon$.

To close out the case $q_1(\ell) < \ell$, I show that either $q_1 < q_2 < 1$ or $\beta(\ell_y^{\min})\, q + y$ does not intersect F twice. To see why, note that if $q_2(\ell) = 1$ then either $y = 1 - 2\pi$ and $\ell = 0$, or $y \in (1 - 2\pi, 1)$ and $\ell = \ell_y^{\min}$. In the former case, any the DTU is actually a UTU, and any UTU intersects F twice: otherwise it lies weakly above F on the interval $[0, 1]$ and

[6] There is at most one value of ℓ such that $\beta(\ell)\, q + y$ is tangent to F in $(0, 1]$.

[7] When $L(x) > 1$, $G_y^{\ell}(x) = 1$ rather than following $L(x)$, but since the line $y = 1$ is an upper bound on F as well, the upper truncation does not affect the behavior of $v(x, \ell)$.

[8] Clearly, given the shape of F, $F(q) > L(q)$ in the interval (q_1, q_2).

strictly above F on some measurable subset of $[0,1]$, and could not have the same mean as F, contradicting the construction of UTUs. Thus $q_1(\ell) < 1$, $v(x,\ell)$ is strictly increasing in (q_1, q_2) and is negative at $x = q_1$, and $G_y^{\beta(\ell)}$ is not Bayes-plausible. In the latter case, if $\beta(\ell_y^{\min})\, q + y$ intersects F twice, then the same argument applies and G_y^{ℓ} is not Bayes-plausible.

If $q_1(\ell) \leq \ell$, then $G_y^{\beta(\ell)}$ satisfies Bayes-plausibility. It must be that $G_\ell(q) < F(q) \;\forall\; q \in (0, q_1) \cup (q_1, q_2)$, with equality at q_1 only if $q_1(\ell) = \ell$. Then $v(x,\ell) > 0$ on $(0, q_2)$, and since $v(x,\ell)$ is strictly decreasing on $(q_2, 1)$ with $v(1,\ell) = 0$, it must be that $v(q) > 0 \;\forall\; q \in (0,1)$. However, I now prove that if $y \leq 1 - 2\pi$, then $G_y^{\beta(\ell)}$ cannot have minimal slope among all Bayes-plausible DTUs with intercept y. Towards proving this claim, I first show that as $\ell \to 0$, it cannot be that $q_1(\ell) \leq \ell$. Assume that for some ℓ_i, L intersects F twice (so that q_1 and q_2 are distinct and well-defined) and $q_1(\ell_i) \leq \ell_i$. Then, for $\ell \in [0, \ell_i]$, the function $\beta(\ell)\, q + y$ will intersect F twice. If $y > 0$, then because $F(0) = 0$ there is $\varepsilon > 0$ so that $\beta(\ell)\, q + y$ lies strictly above F in $[0, \varepsilon)$ for any valid choice of ℓ; thus $q_1(\ell) > \varepsilon$. If instead $y = 0$, then because $f(0) < 1 - 2\pi$, it must be that for any ℓ, there is $\varepsilon > 0$ small enough that $F(\varepsilon) < (1 - 2\pi)\, \varepsilon \leq \beta(\ell)\, \varepsilon$ by convexity of F. Thus it is again true that $q_1(\ell) > \varepsilon$. In either case, taking $\ell < \varepsilon$[9] ensures that $\ell < q_1(\ell)$. To complete the proof, note that $\beta(\ell)\, \ell + y \leq F(\ell)$ is a necessary condition for $q_1(\ell) \leq \ell$. Since $\beta(\ell)\, \ell + y$ is continuous in ℓ, I can use the result above about $\ell \to 0$ to apply the Intermediate Value Theorem and find a value of $\ell \in (0, \pi]$ where $\beta(\ell)\, \ell + y = F(\ell)$ but $\ell - \varepsilon < q_1(\ell - \varepsilon)$ for any $\varepsilon > 0$ sufficiently small. Furthermore, $G_y^{\beta(\ell-\varepsilon)}$ is Bayes-plausible for ε sufficiently small. When $q_1(\ell) = \ell$, it must be that $v(q_1(\ell), \ell) > 0$ since $G_y^{\ell}(q) < F(q) \;\forall\; q \in (0, \ell)$. By continuity of $v(q_1(\ell), \ell)$ in ℓ, it must be that $v(q_1(\ell - \varepsilon), \ell - \varepsilon) > 0$ if ε is sufficiently small. Since $\ell - \varepsilon < q_1(\ell - \varepsilon)$, Equation (A.1) is a necessary and sufficient condition for Bayes-plausibility of $G_y^{\beta(\ell-\varepsilon)}$, and therefore $G_y^{\beta(\ell-\varepsilon)}$ is Bayes-plausible and has a smaller slope than G_y^{ℓ}.

Having established sufficient conditions for when Bayes-plausibility is satisfied, I can now use them to obtain the desired characterization of the slope-minimizing lower truncation length ℓ_y^{sm}. I begin with the case $y \in [0, 1 - 2\pi]$ and show that ℓ_y^{sm} satisfies $v(q_1(\ell_y^{sm}), \ell_y^{sm}) = 0$. When $y \in [0, 1 - 2\pi]$, the lowest permissible slope for a DTU is $(1 - y)^2/(2\pi)$, the slope of the UTU with intercept y. Therefore the line $L(q) = q\,(1 - y)^2/(2\pi) + y$ must intersect F twice in $(0, 1]$. Furthermore, the line $L(q) = q\,(1 - y)/\pi + y$ corresponds to the maximum permissible slope for a DTU, and thus must lie above F for the mean of that DTU to equal the mean of F. Therefore by continuity of $\beta(\ell)$ in ℓ and continuity of f, there exists a value $\ell_t \in (0, \pi)$ where the line $L(q) = \beta(\ell)\, q + y$ is tangent to F. The point of tangency must be interior, as $\beta(\ell_t) \cdot 1 + y = 1$ only if $\ell_t = 0$, in which case the line $\beta(\ell_t)\, q + y$ forms part of a UTU and (as argued above) cannot be tangent to F. Therefore, for $\varepsilon > 0$ sufficiently small the line $\beta(\ell_t - \varepsilon)\, q + y$ intersects F twice, and both intersections are bounded strictly below 1. As argued when showing that $q_1(\ell) \leq \ell$ implies Bayes-plausibility of $G_y^{\beta(\ell)}$, the constraint in Equation (A.1) does not bind for $G_y^{\ell_t}$, so it does not bind for $G_y^{\ell_t - \varepsilon}$, and the latter DTU is therefore Bayes-plausible. Thus the y-optimal DTU G_y^{sm} cannot be tangent to F and must intersect F twice in $(0, 1]$. Since $y \in [0, 1 - 2\pi]$, as shown for the case $q_1(\ell) \leq \ell$ it cannot be that $q_1(\ell_y^{sm}) \leq \ell_y^{sm}$. Therefore $q_1(\ell) > \ell$ and the necessary and sufficient condition for

[9] Of course, this choice may not be valid for $y > 1 - 2\pi$, since the lower bound on the set of valid ℓ is strictly above $\ell = 0$; if so, I cannot rule out that $q_1(\ell) \leq \ell$ for the minimum permissible ℓ.

Bayes-plausibility in Equation (A.1) applies. To show that it holds with equality, consider the UTU corresponding to $\ell = 0$. It is not Bayes-plausible[10] and intersects F twice, so it must be that $v(q_1(0), 0) < 0$. Because $v(q_1(\ell), \ell)$ is a continuous function of ℓ that takes both positive and negative values for $\ell \in [0, \pi]$, the Intermediate Value Theorem implies that there is a well-defined minimum value of ℓ, which I call ℓ_m, for which $v(q_1(\ell_m), \ell_m) = 0$. Since $v(q_1(\ell), \ell) < 0$ for any $\ell < \ell_m$, and I have shown that $v(q_1(\ell_y^{sm}), \ell_y^{sm}) \geq 0$, it must therefore be that $\ell_y^{sm} = \ell_m$.

To complete the proof of the lemma, I show that if $y \in (1 - 2\pi, 1)$, then either $\ell_y^{sm} = \ell_y^{\min}$ or $v(q_1(\ell_y^{sm}), \ell_y^{sm}) = 0$. Assume that $\beta(\ell_y^{\min})\, q + y$ intersects F twice; otherwise clearly $G_y^{\ell_y^{\min}}$ is Bayes-plausible and $\ell_y^{sm} = \ell_y^{\min}$. Assume also that the smallest ℓ for which $v(q_1(\ell), \ell) = 0$, which I label ℓ_y^0, satisfies $\ell_y^0 > \ell_y^{\min}$; otherwise clearly $G_y^{\ell_y^0}$ is both Bayes-plausible and slope-minimizing, so again $\ell_y^{sm} = \ell_y^{\min}$ (if no ℓ satisfying $v(q_1(\ell), \ell) = 0$ exists, I let $\ell_y^0 = \pi$, and the argument still holds). If $\ell_y^{sm} \in (\ell_y^{\min}, \ell_y^0)$, then it must be that $\beta(\ell_y^{sm})\, q + y$ intersects F twice, because $\beta(\ell_y^0)\, q + y$ does. By the definition of ℓ_y^0, $v(q_1(\ell_y^*), \ell_y^{sm}) \neq 0$. Clearly that expression cannot be strictly positive, or by continuity there would be $\varepsilon > 0$ small enough so that $\ell_y^{sm} - \varepsilon$ is both a valid choice of ℓ (i.e., greater than $\ell_y^{\min}$) and generates a Bayes-plausible DTU. It must therefore be strictly negative, which means that $q_1(\ell_y^{sm}) \leq \ell_y^{sm}$; otherwise G_y^{sm} would not be Bayes-plausible. But then the proof that $q_1(\ell) \leq \ell$ cannot occur for small ℓ implies that there is $\varepsilon > 0$ small enough so that $\ell_y^{sm} - \varepsilon > \ell_y^{\min}$ and $G_y^{\ell_y^{sm} - \varepsilon}$ is Bayes-plausible, which contradicts the slope-minimizing property of ℓ_y^{sm} (the caveat for $y > 1 - 2\pi$ does not apply, since we have already covered and ruled out the case $\ell_y^{sm} = \ell_y^{\min}$). Thus it cannot be true that $\ell_y^{sm} \in (\ell_y^{\min}, \ell_y^0)$, so it must be that either $\ell_y^{sm} = \ell_y^{\min}$ or $\ell_y^{sm} = \ell_y^0$; the latter implies the desired condition $v(q_1(\ell_y^{sm}), \ell_y^{sm}) = 0$. □

A.2.4 Overall-Optimal DTUs

The simplified integral constraint in Lemma 11 can be used as a key step in deriving the continuity of ℓ_y^* in y, and thus in providing sufficient conditions for the existence of an overall-optimal DTU in Lemma 12. As an immediate corollary, though, it allows a characterization of the overall-optimal DTU when r^* is small:

Corollary 3. *Let $\beta(0)$ be the slope of the 0-optimal double-truncated uniform distribution $G_0^{\beta(0)}$, and let $q_i(\beta(0), 0)$ be the smallest $q \in (0, 1)$ that satisfies $\beta(0)\, q = F(q)$.*[11] *If $r^* \leq q_i(\beta(0), 0)$, then $G_0^{\beta(0)}$ is uniquely optimal among all double-truncated uniform distributions.*

Proof. The proof is by contradiction, and resembles the geometric proof of Proposition 1 for the binary-state setting. Fix r^* and assume some other DTU G does weakly better than $G_0^{\beta(0)}$ for Sender. It must therefore have a smaller slope than G_0^*: the intercept of G is larger than that of $G_0^{\beta(0)}$, and G must intersect the horizontal line $y = 1$ at a larger value of q than $G_0^{\beta(0)}$ or the concavification of G would be everywhere above that of $G_0^{\beta(0)}$. Because of its

[10] Any UTU with $y = 0$ has an atom at 0 while F does not. If $y = 0$, the restriction that $f(0) < 1/(2\pi)$ ensures that the UTU is not Bayes-plausible, since there is $\varepsilon > 0$ such that the UTU places more mass in the interval $[0, \varepsilon]$ than does F.

[11] The existence of $q_i(\beta(0), 0)$ is guaranteed by the proof of Lemma 11.

larger slope, $G_0^{\beta(0)}$ upper-bounds G after r^* (where G lies weakly below $G_0^{\beta(0)}$) and thus

$$\begin{aligned}
&\int_{r^*}^1 G(q)\,dq < \int_{r^*}^1 G_0^{\beta 0}(q)\,dq \\
&\Rightarrow \int_{q_i(\beta(0),0)}^1 G(q)\,dq < \int_{q_i(\beta(0),0)}^1 G_0^{\beta(0)}(q)\,dq \\
&\Rightarrow \int_0^{q_i(\beta(0),0)} G(q)\,dq > \int_0^{q_i(\beta(0),0)} G_0^{\beta(0)}(q)\,dq = \int_0^{q_i(\beta(0),0)} F(q)\,dq.
\end{aligned}$$

The inequality in the first line is strict because $F(q_1(\beta(0),0)) < 1$, so r^* is not in the upper-truncated region of G and there is some strict difference between $G_0^{\beta(0)}$ and G captured in the integral. The first implication follows from the bound on r^*. The inequality in the third line is because all DTUs have equal means, so

$$\begin{aligned}
1-\pi &= \int_0^1 G(q)\,dq = \int_0^{q_i(\beta(0),0)} G(q)\,dq + \int_{q_i(\beta(0),0)}^1 G(q)\,dq \\
&= \int_0^1 G_0^{\beta(0)}(q)\,dq = \int_0^{q_i(\beta(0),0)} G_0^{\beta(0)}(q)\,dq + \int_{q_i(\beta(0),0)}^1 G_0^{\beta(0)}(q)\,dq.
\end{aligned}$$

The equality in the third line is by Lemma 11, since by Lemma 10 the DTU $G_0^{\beta(0)}$ has minimal slope among Bayes-plausible DTUs with intercept 0. □

Using the characterization of Lemma 11, I now prove a sufficient condition on F for U_{r^*}, the set of utilities attained by y-optimal DTUs, to be compact, and thus for Sender to have a well-defined overall-optimal DTU:

Lemma 12. *Let $r^* \in [\pi, 1]$ and $f(1) > 0$. Then Sender's maximum utility over all double-truncated uniform distributions is well-defined, and is attained by a double-truncated uniform distribution G^*.*

Proof. I first show that the optimal lower truncation length ℓ_y^* is continuous in y at any $y \in [0, 1)$. Given the restriction on r^*, Sender's utility from a y-optimal DTU $G_y^{\beta(y)}$ is given by $1 - G_y^{\beta(y)} = 1 - (\beta(y), y)\, r^* + y)$. Thus continuity of ℓ_y^* in y ensures that Sender's maximum utility over DTUs with intercept y is continuous in y. I can then provide sufficient conditions for the intercept of a potential overall-optimal DTU to lie in a compact set. The continuity condition implies that U_{r^*} is compact, so that it contains its closure. Therefore there is some DTU that attains Sender's supremum utility over all DTUs.

To show continuity, I first work with $y \in [0, 1 - 2\pi)$, where the argument is most straightforward. Since in that range $v(q_1(\ell_y^*, y), \ell_y^*) = 0$ by Lemma 11, and the proof of that lemma shows that ℓ_y^* is the minimal ℓ where the property holds, I can apply the Implicit Function Theorem to write ℓ_y^* as a continuous function of y.

When $y \in (1 - 2\pi, 1)$, then Lemma 11 implies that either $v(q_1(\ell_y^*, y), \ell_y^*) = 0$ or $\ell_y^* = \ell_y^{\min}$. In particular, ℓ_y^* is either the minimum permissible ℓ or, if that choice does not deliver a Bayes-plausible DTU, the minimum ℓ satisfying $v(q_1(\ell, y), \ell) = 0$. Because both $\ell_y^{\min}$ and

the minimal ℓ satisfying $v(q_1(\ell, y), \ell) = 0$ are continuous in y, the minimum over those two choices is also continuous in y. Thus ℓ_y^* is continuous in y for $y \in (1-2\pi, 1)$.

All that remains is to show that ℓ_y^* is continuous in y at $y = 1-2\pi$. The continuity of $\ell_y^{\min}$ in y ensures that the function

$$u(y) = \beta(\ell_y^{\min})\, \ell_y^{\min} + y - F(\ell_y^{\min})$$

is also continuous in y. Because $u(y) > 0$ for any $y \in [0, 1-2\pi]$, as shown in the proof of why $q_1(\ell) > \ell$ for small enough ℓ, it must be that for $\delta > 0$ sufficiently small, $u(y') > 0$ for any $y' \in (1-2\pi, 1-2\pi+\delta)$. Since the line $\beta(0, 1-2\pi)\, q + (1-2\pi)$ intersects F twice in $(0, 1]$, it must therefore be that for $\delta > 0$ sufficiently small and $y' \in (1-2\pi, 1-2\pi+\delta)$, so do the lines $\beta(\ell_{y'}^{\min}, 1-2\pi)\, q + (1-2\pi)$, $\beta(\ell_{y'}^{\min}, y')\, q + (1-2\pi)$, and $\beta(\ell_{y'}^{\min}, y')\, q + y'$. Because the last intersects F twice in $(0, 1]$, and both intersections occur at values $q > \ell_{y'}^{\min}$, the proof of Lemma 11 shows that $v(q_1(\ell_{y'}^*, 0), \ell_{y'}^*) = 0$ and $\ell_{y'}^*$ is the minimal value of ℓ such that this property holds. Therefore, by the continuity of the minimal value of ℓ satisfying this equation, ℓ^* is continuous in y at $y = 1-2\pi$.

Having shown continuity of ℓ_y^* in y, I use the second part of the lemma statement to show that the set of possibly overall-optimal DTU intercepts is compact. Given that $f(1) > 0$, there must be $\bar{y} \in (0, 1)$ such that $1 - \bar{y} > f(1)$. Then for any intercept $y \geq \bar{y}$, the DTU with minimal permissible slope lies above F on $(0, 1)$, and is therefore Bayes-plausible. Since any DTU with intercept $y > \bar{y}$ surely lies above the slope-minimal DTU with intercept $\bar{y}$ for all $q \in [\pi, 1]$, no DTU with intercept in $(\bar{y}, 1)$ can be optimal among all DTUs. Thus the intercept of the overall-optimal DTU lies in the compact set $[0, \bar{y}]$. □

Note that only the last step of the proof relies on $f(1) > 0$; if this condition is violated, then (as in the statement of Proposition 3 in the text) it may be that no DTU attains Sender's supremum utility, but there exists a limiting sequence of DTUs converging to that value so no distribution delivers strictly higher utility than all DTUs.

A.2.5 Optimal Posterior Distributions

Having established properties of overall-optimal DTUs, I can now jointly prove the optimal distribution portions of Propositions 2 and 3:

Proof. Let H be a candidate optimal distribution of posterior means. I approximate $\bar{H}$, the concavification of H, by a tangent at r^*, which I call $L(q)$; let $L(0) = y_L \in [0, 1)$ be its intercept. Consider the y_L-optimal DTU $G_{y_L}^{\beta(y_L)}$. In order for H to do at least as well for Sender as $G_{y_L}^{\beta(y)}$, by Lemma 3 it must be that

$$1 - H(r^*) \geq 1 - \bar{H}(r^*) \geq 1 - G_{y_L}^{\beta(y_L)}(r^*).$$

Thus L must have a weakly smaller slope than $G_{y_L}^{\beta(y_L)}$; otherwise $L(r^*) > G^{\beta}(y_L)_{y_L}(r^*)$ and the above inequality is violated.

If $y_L \in [0, 1-2\pi]$, then for any slope $\beta \in ((1-y_L)^2/(2\pi), \beta(y_L)]$ there is a DTU with that slope and intercept y_L. If instead $y_L \in (1-2\pi, 1)$, then for any slope $\beta \in [1-y_L, \beta(y_L)]$ there is a DTU with that slope and intercept y_L. In the first case, the slope of L cannot lie

below that interval or it would have a weakly smaller slope than the UTU with intercept y_L; then the argument of Proposition 1 applies and H is not Bayes-plausible. In the second case, L must have a slope weakly greater than the lowest-slope DTU with intercept y_L, or it would fail to pass through $(1,1)$, and therefore so would $\bar{H}$ and H. Thus there is a DTU G^L with the same slope as L.

Let $r^* \in (0, q_i(\beta(0),0)]$. By Corollary 3, if $G^L \neq G_0^{\beta(0)}$, then because $G^L(r^*) \leq G_0^{\beta(0)}(r^*)$, G_L is not Bayes-plausible. If instead $r^* \in [\pi, 1)$ and G^L has a strictly smaller slope than $G_{y_L}^{\beta(y_L)}$, then by Lemma 10, G^L is not Bayes-plausible.

In either case, given that G^L violates Bayes-plausibility, H must violate it as well. Because G^L upper-bounds $\bar{H}$ beyond ℓ, it must be that

$$\int_q^1 H(t)dt \leq \int_q^1 \bar{H}(t)dt \leq \int_q^1 G^L(t)dt$$

for any $q \in [\ell, 1]$. Since G^L violates Bayes-plausibility, there is some $q_v \in [0,1]$ where

$$\int_0^{q_v} G^L(t)dt > \int_0^{q_v} F(t)dt,$$

and since the left-hand side equals 0 for any $q \in [0, \ell)$, it must be that $q_v \in [\ell, 1]$. Then because H and G^L have the same mean,

$$\begin{aligned}
&\int_0^1 H(t)dt = \int_0^1 G^L(t)dt = 1 - \pi \\
&\Rightarrow \int_0^{q_v} H(t)dt + \int_{q_v}^1 H(t)dt = \int_0^{q_v} H(t)dt + \int_{q_v}^1 H(t)dt \\
&\Rightarrow \int_0^{q_v} H(t)dt \geq \int_0^{q_v} G^L(t)dt > \int_0^{q_v} F(t)dt,
\end{aligned}$$

where the third line follows from the earlier upper bound on the integral of H. Therefore H violates Bayes-plausibility and is not a valid distribution.

If G^L has the same slope as $G_{y_L}^{\beta(y_L)}$, then by construction H gives Sender the same utility as $G_{y_L}^{\beta(y_L)}$. Thus if there is a DTU that delivers Sender a strictly higher utility than $G_{y_L}^{\beta(y_L)}$, then clearly H is not optimal overall. If there is no such DTU, then $G_{y_L}^{\beta(y_L)}$ is optimal among all DTUs and H also attains Sender's maxmin utility. □

Finally, I prove the unique concavification portion of Proposition 2:

Proof. By the proof for optimal distributions above, the slope of $\bar{H}$ at r^* equals that of $G_0^{\beta(0)}$. Because $\bar{G}_0^{\beta(0)}$ does not have a kink at ℓ_0^*, it upper-bounds H on the whole interval $[0,1]$ instead of just on $[\ell, 1]$ as in that proof. If $H < G_0^{\beta(0)}$ on any measurable subset of $[r^*, 1]$ the proof of Corollary 3 shows that H violates the Bayes-plausibility integral constraint at $q_i(\beta(0), 0)$.

If $\ell_0^* \leq r^*$, it is therefore true that $\bar{H} = G_0^{\beta(0)} = \bar{G}_0^{\beta(0)}$ on $[r^*, 1]$. Furthermore, $\bar{G}_0^{\beta(0)}$ upper-bounds $\bar{H}$ on $[0, r^*]$ and $\bar{G}_0^{\beta(0)}(0) = \bar{H}(0) = 0$. Because $\bar{G}_0^{\beta(0)}$ is linear on $[0, r^*]$ (i.e., it

has no kink at ℓ_0^*) there is no smaller concave function that takes the same values at $q = 0$ and $q = r^*$; thus $\bar{H} = \bar{G}_0^{\beta(0)}$ on $[0, r^*]$ as well.

If instead $\ell_0^* > r^*$, then it is now the case that $\bar{H} = G_0^{\beta(0)} = \bar{G}_0^{\beta(0)}$ on $[\ell_0^*, 0]$, since that is the range where the latter equality holds. However, the upper-bounding relationship still holds on $[0, \ell_0^*]$, and thus the argument above still applies and $\bar{H} = \bar{G}_0^{\beta(0)}$ on $[0, \ell_0^*]$. □

A.2.6 The Finite-State Case

Note that the proofs and results of Lemmas 9 and 10 go through unchanged. Thus I can in fact easily prove an analogue for Proposition 3 by following the proof in Appendix A.2.5. In particular, a candidate optimal distribution can be approximated by a DTU G^L. If G^L has a strictly smaller slope than the y-optimal DTU with the same intercept y, then by assuming $r^* \in [\pi, 1]$ and following Lemma 10, it cannot be Bayes-plausible. Given that G^L is not Bayes-plausible, neither is the candidate optimal distribution. Since there is a y-optimal DTU for any possible y, no distribution can give Sender strictly higher utility than all DTUs.

To obtain a tighter characterization of which DTU is optimal in this setting, I can prove an analogue of Lemma 11, showing where the integral constraint binds for the y-optimal DTU:

Lemma 13. *Let $y = 0$ or $r^* \in [\pi, 1]$ so that the slope-minimizing Bayes-plausible DTU is also Sender's optimal DTU. Fix the value of y, and let q_i be the location of the ith interior atom of the prior F. Then either the optimal lower truncation length ℓ_y^* is equal to the minimum lower truncation length ℓ_y^{min}, or for at least one $i \in \{1, ..., N-2\}$ it is true that*

$$\lim_{x \to q_i^-} \int_0^x F(q)\, dq - \int_0^x G_y^\beta = 0.$$

Proof. For completeness, define $q_0 = 0$.

The proof is algorithmic; the algorithm for finding the optimal DTU is as follows.

1. Initialize β as the minimal feasible slope for a DTU with intecept y.
2. For $i \in 1, ..., N-2$:
   (a) Check whether the integral constraint is satisfied at the left limit of q_i. That is, whether
   $$\lim_{x \to q_i^-} \int_0^x F(q)\, dq - \int_0^x G_y^\beta \geq 0.$$
   (b) If the constraint is satisfied, exit.
   (c) Else, increase β until the constraint binds exactly.[12]
3. Return $\beta = \beta(y)$, the y-optimal slope.

[12] By the proof of Lemma 9, the difference of integrals is continuous and monotonically increasing in β, so there will be exactly one value where the constraint binds.

In step (b), if the integral constraint is satisfied at the left limit of q_i, it must be satisfied everywhere in $[q_{i-1}, q_i)$, since in that interval F is constant but G_y^β is weakly decreasing. The algorithm first finds the minimal value β where G_y^β satisfies the integral constraint in $[0, q_1)$, then proceeds across subsequent intervals, increasing β if necessary to ensure the integral constraint is satisfied. Finally, the integral constraint is automatically satisfied in $[q_{N-2}, 1]$ because G_y^β has the appropriate mean.

Thus either the initial value of β satisfies the integral constraint for all i, in which case that minimal feasible slope is y-optimal, or the constraint is satisfied for at least one i, giving the result in the lemma. □

As in the continuous-state case, the integral constraint binds at a finite and possibly empty set of points for each y-optimal DTU. In the continuous-state case, this set was guaranteed to be nonempty for all $y \in [0, 1 - 2\pi]$; however if $F(0) > 0$ in this finite-support setting, the set may be empty for even $G_y^\beta(0)$, the 0-optimal DTU. In the case where there does exist a minimal q_i—call it $q_{\min}$—where the integral constraint binds for $G_0^{\beta(0)}$ (a fact which depends on the specification of F), there is a natural analogue of Corollary 3: if $r^* \in [0, q_{\min}]$, then $G_0^{\beta(0)}$ is Sender's overall-optimal DTU. The proof exactly parallels that of the original corollary. Indeed, the analogue of Proposition 2 also follows, since the proof of optimality in Appendix A.2.5 then goes through in the same way.

A.3 A Numerical Approach to the Continuous-State Setting

A.3.1 Summary of Numerical Results

Propositions 2 and 3 leave open the optimal distribution of posterior means when the mean Receiver type r^* lies in $\pi \in (q_i(\beta(0), 0), \pi)$. While the value $q_i(\beta(0), 0)$ is well-defined for a given prior distribution, a closed-form solution may not exist. However, fixing a prior distribution, I can use a two-step solution algorithm to numerically compute $q_i(\beta(0), 0)$ and show qualitatively how the size of the intermediate interval changes with various properties of the prior distribution.

Informally, given a prior F with mean π, the first step is to find the 0-optimal slope $\beta(0)$. To check Bayes-plausibility, I use the simplified integral constraint from Lemma 11. Starting with the minimum β, I increase β only if the constraint is violated and stop when it binds. The second step checks for intersections between the 0-optimal DTU and F; by definition the smallest interior intersection is $q_i(\beta(0), 0)$. A full formal description is in Appendix A.3.2.

I briefly discuss some intuition for the results of the numerical computation below. A reader interested in further detail may refer to Appendix A.3.3 for thorough figures showing the output of the algorithm at various parameter values, or to Appendix A.3.4 for a detailed exposition of those figures. Throughout this section μ and σ refer to the mean and standard deviation of the generating normal distribution $N(\mu, \sigma^2)$ while π refers to the mean of the prior F, i.e., $N(\mu, \sigma^2)$ truncated in $[0, 1]$.

For fixed μ, the 0-optimal slope $\beta(0)$ and the 0-optimal lower truncation length $\ell(0)$ are decreasing in σ. A smaller slope is better for Sender, but may be ruled out by the integral constraint; increasing σ means the prior cdf increases less steeply, so the integral constraint allows Sender's chosen distribution to increase less steeply as well.

For $\mu < 1/2$ (the midpoint of the truncation interval), the key value $q_i(\beta(0), 0)$ has an inverse-U-shaped relationship with σ, while for $\mu > 1/2$ it is increasing in σ. Increasing σ lowers the slope of the 0-optimal DTU, which would decrease $q_i(\beta(0), 0)$ if the shape of the prior were unchanged. However, holding fixed a DTU's slope, increasing σ spreads out the prior mass and increases $q_i(\beta(0), 0)$. For small μ, either of these effects can dominate. For large μ, most of the prior mass is far enough away from the origin that the second effect dominates.

The prior mean π always lies strictly below $q_i(\beta(0), 0)$ for σ small enough, and increases with σ when $\mu < 1/2$ but decreases with σ when $\mu > 1/2$. This behavior is a known property of the truncated normal distribution; in my setting, it implies that there is no gap between Proposition 2 and Proposition 3 whenever $\mu > 1/2$. When $\mu < 1/2$, it implies that eventually $q_i(\beta(0), 0) < \pi$, producing a gap between the results.

A.3.2 A Detailed Algorithm for Computing the Smallest Interior Intersection

In this section I describe in detail the algorithm for computing $q_i(\beta(0), 0)$, as well as some notes on its key steps and the details of implementation.

1. Compute the 0-optimal slope $\beta(0)$:

   (a) Following Appendix A.2.1, the minimal DTU slope when the intercept is $y = 0$ is $\beta_{\min} = \max\{1/(2\pi), 1\}$. Initialize the slope β at $\beta_{\min}$.

   (b) Define the function
   $$\ell(\beta) = \frac{\sqrt{2\pi\beta - 1}}{\beta},$$
   the lower truncation length of a DTU with slope β and intercept $y = 0$ (as described in Appendix A.2.1).

   (c) Given β, define the function
   $$v(q) = \int_0^q F(x)\,dx - \int_0^x G^{\beta}_{\ell(\beta)}(x)\,dx$$

   (d) Compute $v^* = \min_{q\in(0,1)} v(q)$ using standard numerical optimization.

   (e) If $v^* < 0$, increase β and return to 2(b). Else, return $\beta(0) = \beta$.

2. Find the smallest interior intersection $q_i(\beta(0), 0)$:

   (a) Given $\beta(0)$, solve $\beta(0)\,q = F(q)$ using a standard numerical solver.

   (b) Discard the solution $q = 0$; return the smallest remaining solution as the value of $q_i(\beta(0), 0)$.

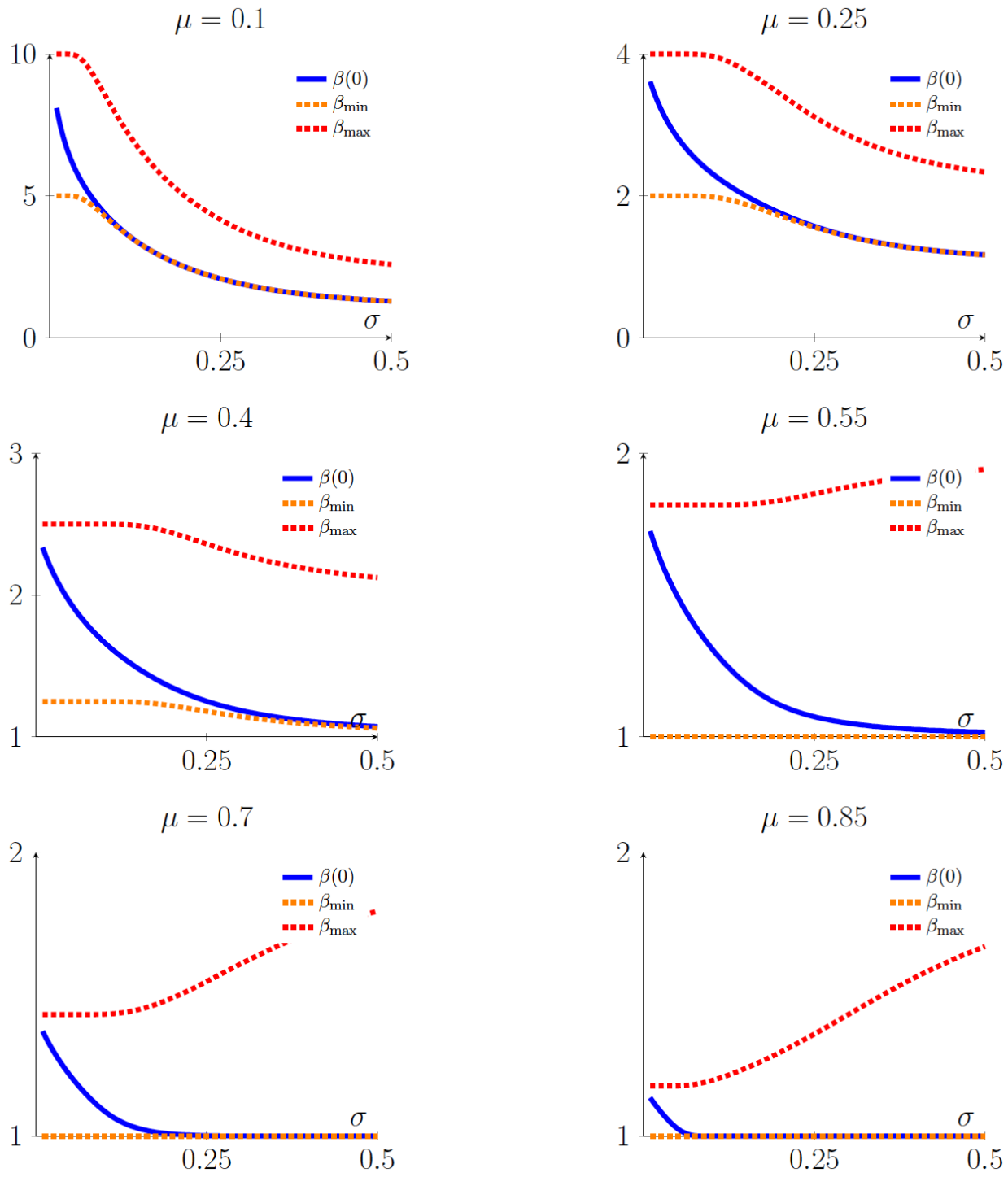
Step 1(e) is guaranteed to terminate because for $\beta = \beta_{\max} = 1/\pi$ the DTU will lie weakly above F everywhere after the lower truncation region $[0, \ell(\beta)]$; following the proof of Lemma 11, if this relationship holds then $G^{\beta}_{\ell(\beta)}$ is Bayes-plausible. Step 2(a) is guaranteed to find an non-zero solution, a result of the same proof.

In practice, step 2 was computed by defining the minimum and maximum values of β given π (as described in Appendix A.2.1) and considering a 1,000-point grid over that interval. Then, since an analytic expression for the truncated normal does not exist, the integral in step 2(c) was computed numerically over a 10,000-point grid covering $[0, 1]$. The procedure was repeated for prior distributions constructed using $\mu = (0.01, 0.02, ...0.99)$ and $\sigma = (0.01, 0.015, ...0.50)$ for the generating normal distribution. Larger values of σ^2 produced errors; even the maximum precision of the numerical integral was not high enough to accurately complete steps 1(d) and 1(e).

A.3.3 Plotting Zero-Optimal DTUs

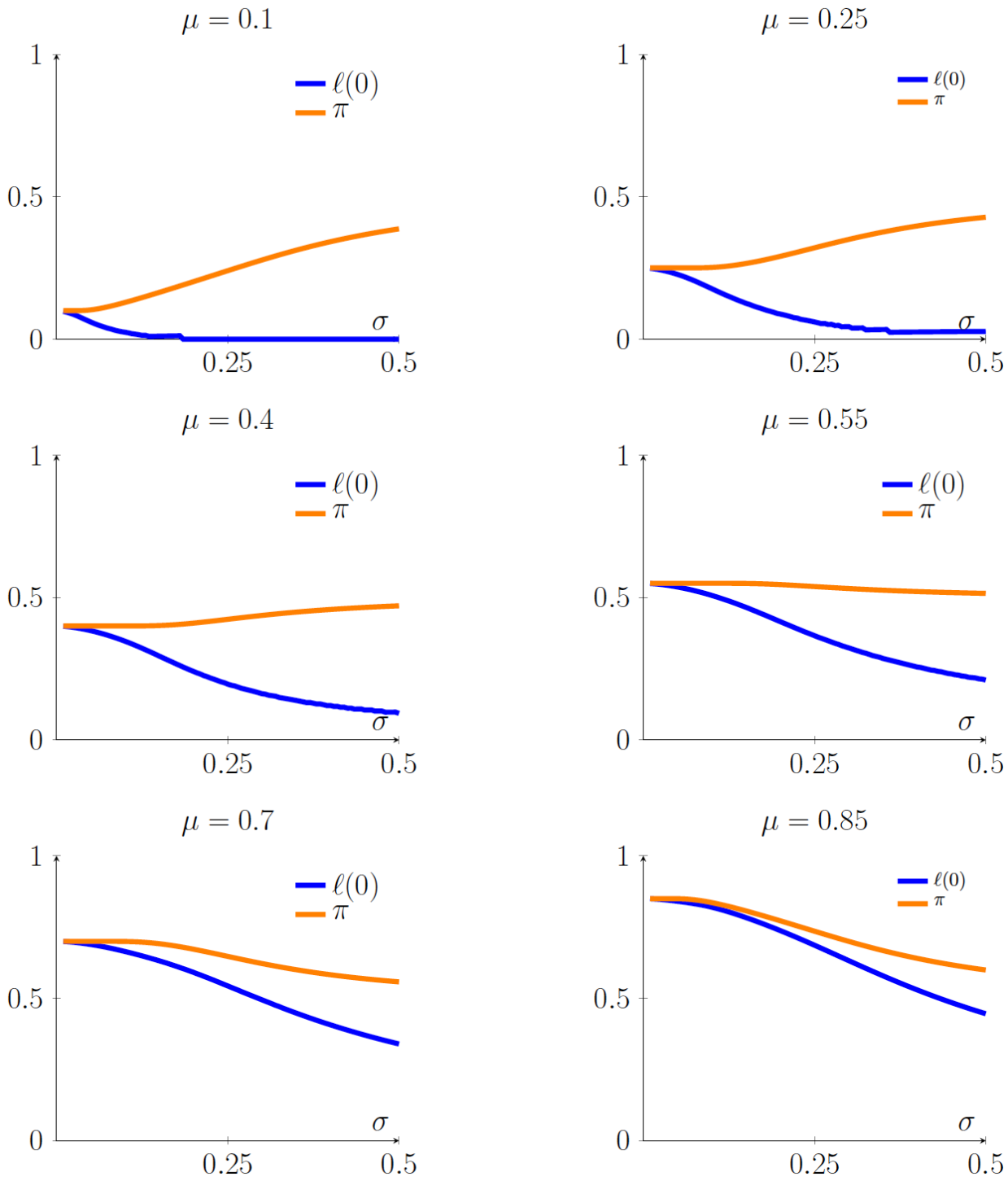
Below I include detailed plots of the 0-optimal slope $\beta(0)$, the 0-optimal lower truncation length $\ell(0)$, and the value $q_i(\beta(0), 0)$ used in Proposition 2. In each figure, the prior is a truncation to $[0, 1]$ of a $N(\mu, \sigma^2)$ distribution. Each panel fixes a value $\mu \in \{0.1, 0.25, 0.4, 0.55, 0.7, 0.85\}$ and shows the relevant values as functions of the standard deviation σ.

Figure A.1: Numerically Computed Values of $\beta(0)$



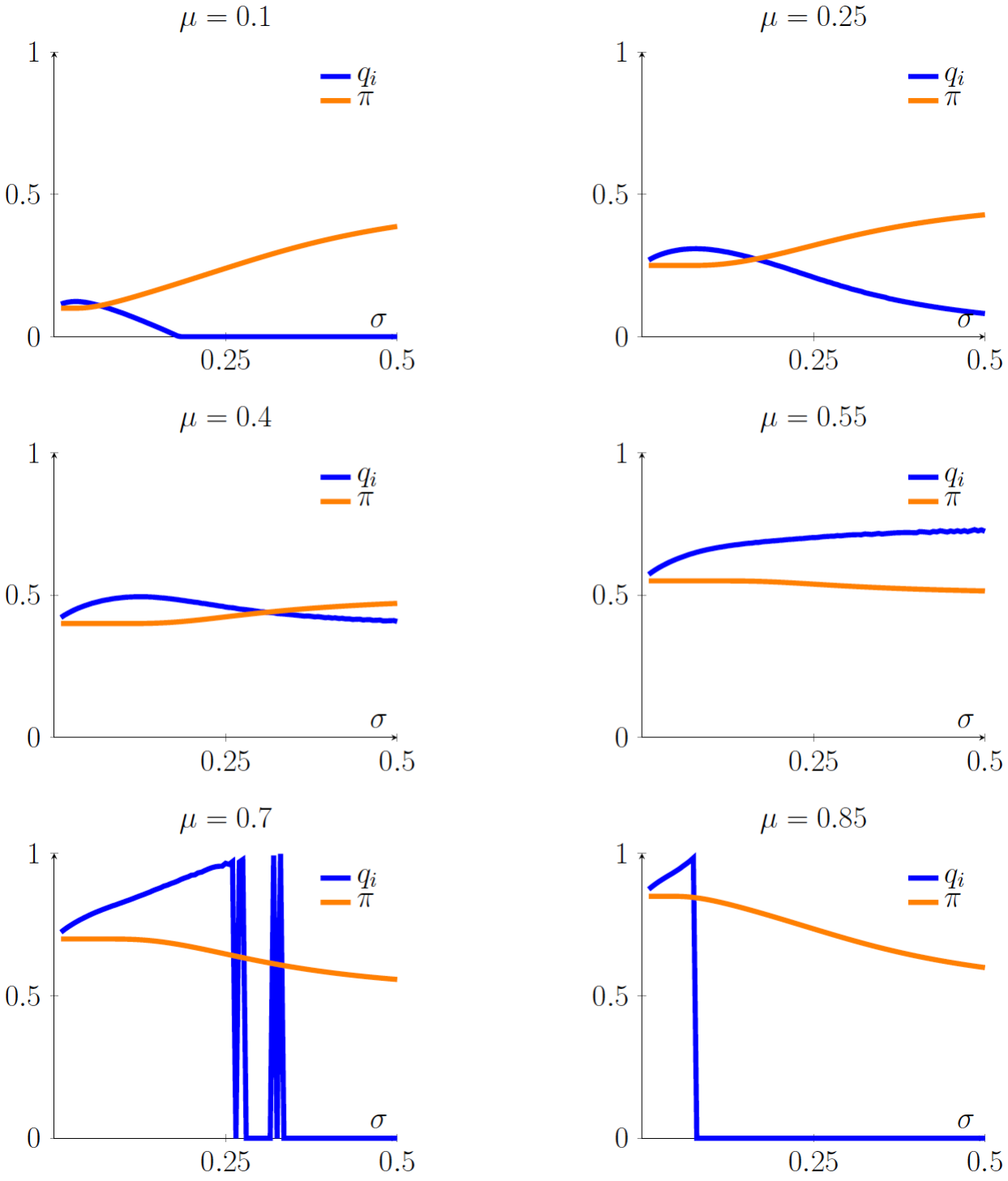
For fixed μ, values of $\beta(0)$ and the upper and lower bounds on β as functions of σ.

Figure A.2: Numerically Computed Values of $\ell(0)$



For fixed μ, values of $\ell(0)$ and the upper bound $\ell_{\max} = \pi$ as functions of σ.

Figure A.3: Numerically Computed Values of $q_i(\beta(0), 0)$



For fixed μ, values of $q_i(\beta(0), 0)$ and π as functions of σ.

A.3.4 Discussion of Numerical Results

Figure A.1 shows clear patterns in the 0-optimal choice of β, $\beta(0)$, across prior parameter values. Fixing μ, $\beta(0)$ is monotonically decreasing in σ with a roughly exponential shape. As μ increases, the range of $\beta(0)$ decreases. Note that for σ large, the assumption $f'(0) < 1 - 2\pi$ used in Section 1.5 to rule out the binary-state solution may be violated. In this case, Proposition 2 does not apply as $\beta(0) = \beta_{\min}$ may be a feasible solution. This case occurs in panels 1 and 2 of the figure. A similar phenomenon occurs when μ is large, so that the prior F is concave (as described in Section 1.5.1), and appears in panels 5 and 6 of the figure.

The 0-optimal lower truncation length $\ell(0)$ behaves as expected given the results for $\beta(0)$, and is shown in Figure A.2. For fixed μ, $\ell(0)$ is monotonically decreasing and appears to have a reverse-S shape (concave and then convex). Increasing μ shifts the curve up and flattens it. As with $\beta(0)$, the results in the first two panels show some instability at high values of σ resulting from violations of the assumption $f'(0) < 1 - 2\pi$. In panels 5 and 6, where $\beta(0) = \beta_{\min}$ at high values of σ, the lower truncation length tracks directly with the prior mean π in order to maintain Bayes-plausibility of the distribution. Figure A.2 also shows the prior mean π of the truncated normal prior. The effects of truncating the normal distribution are well-known; I only note that π increases with σ for $\mu < 1/2$ and decreases with σ for $\mu > 1/2$ because $\mu = 1/2$ is the midpoint of the truncation interval.

The key value of interest is the smallest interior point of intersection $q_i(\beta(0), 0)$ between the 0-optimal DTU and the prior distribution; I abbreviate this value to q_i for the remainder of this discussion. Figure A.3 illustrates some of the complex interactions between q_i and the shape of the prior, as well as some of the difficulties faced in the numerical approach. For small μ, q_i is inverse-U-shaped as a function of σ. As discussed earlier, σ large enough means that a lower truncation region may not be necessary. In this case I set $q_i = 0$ to preserve continuity of q_i in σ and to reflect that the Bayes-plausibility constraint does not bind at any interior intersection (it trivially binds at a posterior mean of 0, and binds at q_i when a lower truncation is necessary). This case appears in panel 1. As μ increases, the curve flattens and moves up, as seen in panels 2 and 3. For $\mu > 1/2$, q_i becomes monotonically increasing in σ, as shown in panel 4. In panels 5 and 6, because F is convex on $[\varepsilon, 1]$ for small $\varepsilon \geq 0$, all interior intersections between the 0-optimal DTU and F are either close to 0 or close to 1. The numerical algorithm thus becomes unstable and alternates between these two regions (as seen in panel 5) or chooses the default solution of $q_i = 0$ (as seen in panel 6). Nevertheless, the shape of the curve for small σ suggests that the trend of monotonically increasing q_i when $\mu > 1/2$ is preserved.

With respect to the relationship between q_i and π, which determines the gap between Propositions 2 and 3, the numerical results show that the cutoff of $\mu = 1/2$ is key. For any $\mu < 1/2$, there exists $\bar{\sigma}$ large enough so that $\pi > q_i$ for all $\sigma > \bar{\sigma}$. However, $\bar{\sigma}$ may be so large that the assumption $f'(0) < 1 - 2\pi$ is violated, in which case $\pi < q_i$ for all valid choices of σ. For $\mu > 1/2$, $\pi < q_i$ for all σ; to verify this numerically, I spot-checked values $\sigma \in \{1, 10, 100\}$ and manually debugged the numerical integration. Thus there is only a gap between Propositions 2 and 3 when $\mu < 1/2$ and σ is large enough to exceed $\bar{\sigma}$ but not so large as to violate $f'(0) < 1 - 2\pi$. For example, in panel 1 we can see that for $\sigma \in (0.06, 0.165)$, there is a gap between the two propositions. For $\sigma < 0.06$, $q_i > \pi$ so there is no gap, and for $\sigma > 0.165$, the assumption $f'(0) < 1 - 2\pi$ is violated and the propositions do not apply.

Appendix B

Additional Materials for "Discovery through Trial Balloons"

B.1 Proofs for Benchmark Cases

This section provides proofs of optimal discovery for the benchmark cases of perfectly correlated projects (Proposition 4), complementary projects (Proposition 5), and substitutable projects (Proposition 6).

Proof of Proposition 4.

The equivalence of discovering one project and discovering both follows automatically from perfect correlation. We thus refer to discovering both projects throughout, and compare full discover to no discovery only.

If $\mu_1 < 0$ and $\mu_2 < 0$, the payoff without discovery is 0 for sure, whereas discovery may give a positive payoff (if the realized projects have sufficiently high value).

If $\mu_1 + \mu_2 \geq 0$, then the payoff without discovery is 1 for sure, whereas discovery may give a lower payoff (if the realized projects have sufficiently low value).

If $\mu_i = -\mu < 0 \leq c\mu = \mu_j$ for $c \in [0, 1)$, then the payoff without discovery is w_j for sure, as only project j is approved. Consider discovering project i; then if $c = 1/(2 - \sigma_1/\sigma_2)$, by symmetry of the marginal distribution of v_i there is an equal probability that

$$v_i < \mu_i - (\sigma_i/\sigma_j)\,\mu_j \Rightarrow v_j < \mu_j - \mu_j = 0,$$

so μ_i is no longer approved, and

$$v_j > \mu_i + (\sigma_i/\sigma_j)\,\mu_j \Rightarrow v_i + v_j = (\mu_i + \mu_j) + (1 + \sigma_j/\sigma_i)\,\mu_i > 0.$$

If $w_j = 1/2$, the expected payoff is therefore the same as no discovery.

If instead $w_j < 1/2$, then the expected payoff from discovering project 1 is greater than from no discovery, because the increase in utility from getting 1 instead of w_j exceeds the loss from getting 0 instead of w_j. As $c \to 0$ with μ held fixed, the probability that $v_j < 0$ approaches $1/2$, whereas the probability that $v_i + v_j > 0$ approaches $\mathbb{P}(v_i > 0) = \Phi(\mu/\sigma_1) := p < 1/2$. We therefore have an expected utility of $(1 - 1/2 - p) \cdot w_j + p \cdot 1$ in the limit. When $w_j > 2p/(2p + 1)$, this limit value is less than w_j, and so by continuity there is a cutoff

$\underline{c} \in (0, 1/(2-\sigma_1/\sigma_2))$ such that discovering project 1 is best if and only if $c > \underline{c}$. When $w_j < 2p/(2p+1)$, discovering project 1 is best for all c.

Finally, if $w_j > 1/2$, then the expected payoff from discovering project 1 is less than from no discovery. As $c \to 1$, the probability that $v_j < 0$ approaches $\Phi(-\mu/\sigma_2) := q < 1/2$ whereas the probability that $v_i + v_j > 0$ approaches $1/2$. We therefore have an expected utility of $(1 - 1/2 - q)w_j + 1/2 \cdot 1$ in the limit. When $w_j < 1/(2q+1)$, this limit value exceeds w_j, and so by continuity there is a cutoff $\underline{c} \in (1/(2-\sigma_1/\sigma_2), 1)$ such that discovering project 1 is best if and only if $c > \underline{c}$. When $w_j > 1/(2q+1)$, discovering nothing is best for all c. □

Proof of Proposition 5.

Throughout, let X be a $N(0,1)$ random variable. Note that upon discovering project i and obtaining realization v_i, the agent's posterior belief about v_j is

$$v_j | v_i \sim N\left(\mu_j + \rho \frac{\sigma_j}{\sigma_i}(v_i - \mu_i), (1-\rho^2)\sigma_j^2\right).$$

Thus the probability of obtaining the grand bundle is given by

$$\begin{aligned}
\mathbb{P}(v_i + \mathbb{E}(v_j|v_i) \geq 0) =& \mathbb{P}\left(v_i + \mu_j + \rho\frac{\sigma_j}{\sigma_i}(v_i - \mu_i) \geq 0\right) \\
&= \mathbb{P}\left(v_i \geq \frac{(\rho\sigma_j/\sigma_i)\mu_i - \mu_j}{1 + (\rho\sigma_j/\sigma_i)}\right) \\
&= \mathbb{P}\left(X \geq \left(\frac{(\rho\sigma_j/\sigma_i)\mu_i - \mu_j}{1 + (\rho\sigma_j/\sigma_i)} - \mu_i\right)/\sigma_i\right) \\
&= \mathbb{P}\left(X \geq -\frac{\mu_i + \mu_j}{\sigma_i + \rho\sigma_j}\right).
\end{aligned}$$

When discovering both projects, the probability of obtaining the grand bundle is

$$\mathbb{P}(v_i + v_j \geq 0) = \mathbb{P}\left(X \geq -\frac{\mu_i + \mu_j}{\sqrt{\sigma_i^2 + \sigma_j^2 + 2\rho\sigma_i\sigma_j}}\right),$$

where we note that the sum of two correlated normals is normal with mean and variance as given above (i.e., larger variance but same mean as if they were uncorrelated).

It is clear that whenever $\sigma_i > \sigma_j$, discovering only project i is better than discovering only project j, and vice-versa. This result follows from noting that the only way of attaining the grand bundle is to draw an above-mean realization from the appropriate marginal, and so the higher-variance marginal distribution is the best choice.

Setting $\sigma_i = 1$ (which is without loss because we can then scale the other parameters accordingly), we can also obtain a condition for discovering both projects to be better than discovering only i:

$$\begin{aligned}
-\frac{\mu_i + \mu_j}{1 + \rho\sigma_j} &> -\frac{\mu_i + \mu_j}{\sqrt{1 + \sigma_j^2 + 2\rho\sigma_j}} \\
(1-\rho^2)\sigma_j^2 &> 0.
\end{aligned}$$

This condition holds for all valid ρ, so the absence of restrictions on σ_j, μ_i, or μ_j means that discovering both is in fact better than discovering either individual project.

□

Proof of Proposition 6.

We first show that, when restricted to only discovering a single project, discovering the project i satisfying $\sigma_i/\mu_i < \sigma_j/\mu_j$ is better than discovering the other project. Using the expressions from the proof of Proposition 5, we can see that upon discovering project 1, the probability of having a project approved is

$$\max\left\{\mathbb{P}\left(X \geq -\frac{\mu_1}{\sigma_1}\right), \mathbb{P}\left(X \geq -\frac{\mu_2}{\rho\sigma_2}\right)\right\},$$

where X is a standard normal random variable. Similarly, the probability of having a project approved upon discovering project 2 is

$$\max\left\{\mathbb{P}\left(X \geq -\frac{\mu_1}{\rho\sigma_1}\right), \mathbb{P}\left(X \geq -\frac{\mu_2}{\sigma_2}\right)\right\}.$$

From here we can see that when $\sigma_1 \geq \sigma_2\,\mu_1/(\rho\mu_2)$, both maxima are given by the first probability, and so discovering project 1 is best. When $\sigma_1 \leq \sigma_2\,\rho\mu_1/\mu_2$, both maxima are given by the second probability, and so discovering project 2 is best. In the remaining interval, we compare $\mathbb{P}(X \geq -\mu_1/\sigma_1)$ to $\mathbb{P}(X \geq -\mu_2/\sigma_2)$, giving precisely the cutoff in the proposition.

We now turn to the comparison between discovering both projects and discovering only one. We drop the assumption that $\mu_1 < \mu_2$ and instead without loss of generality we normalize $\sigma_1^2 = 1$ (adjusting μ_1, μ_2, and σ_2^2 as we wish).

If only project 1 is discovered, then there is a level v^* such that for all $v_1 \geq v^*$, the principal will get at least one project approved, and below that point they will get neither project approved. To be able to get project 1 approved, the principal needs $v_1 \geq 0$. To be able to get project 2 approved, she needs $\mathbb{E}[v_2|v_1] \geq 0$, which we can write as

$$\mathbb{E}[v_2|v_1] = \mu_2 + \rho\sigma_2(v_1 - \mu_1) \geq 0 \iff v_1 \geq \frac{\rho\sigma_2\mu_1 - \mu_2}{\rho\sigma_2},$$

where we assume $\rho > 0$ so as not to reverse the inequality when dividing by $\rho\sigma_2$.

If $\mu_2 \leq \rho\sigma_2\mu_1$, then the value of v_1 needed for $\mathbb{E}[v_2|v_1] \geq 0$ is weakly positive, so $v^* = 0$. Thus in this case, when discovering one project, the principal gets at least one project approved if and only if $v_1 \geq 0$. When discovering both, the principal gets at least one project approved with strictly higher probability, because even when $v_1 < 0$ it is still possible to realize $v_2 \geq 0$, and in that case the second project is approved.

Assume instead that $\mu_2 > \rho\sigma_2\mu_1$, so that $v^* < 0$. In this case, if $v_1 \geq 0$ then regardless of whether project 2 is discovered, the principal surely gets project 1 approved. If $v_1 \in [v^*, 0)$, then when the principal does not discover project 2, it is surely approved; if the principal does discover project 2, it may or may not be approved. If $v_1 < v^*$, then when the principal does not discover project 2, neither project is approved; if the principal does discover project 2, it may or may not be approved. Thus we can write the difference between discovering both

projects and discovering project 1 only as $\delta(v_1)$, a function of the value v_1 of the discovered project 1:

$$\delta(v_1) = \begin{cases} \mathbb{P}(v_2 \geq 0 \,|\, v_1) - 0 = 1 - \Phi\big(-\frac{\mu_2 + \rho\,\sigma_2\,(v_1 - \mu_1)}{\sigma_2\sqrt{1-\rho^2}}\big) := 1 - \Phi(\lambda(v_1)), & v_1 < v^* \\ \mathbb{P}(v_2 \geq 0 \,|\, v_1) - 1 = -\Phi(-\frac{\mu_2 + \rho\,\sigma_2\,(v_1 - \mu_1)}{\sigma_2\sqrt{1-\rho^2}}) := -\Phi(\lambda(v_1)), & v_1 \in (v^*, 0) \\ 0 & v_1 \geq 0 \end{cases}$$

We aim to show that, in expectation, $\delta(v_1)$ is positive, so that it is better for the principal to discover both projects than to discover project 1 only.

$$\begin{aligned} & \int_{-\infty}^{\infty} \delta(v)\,dv > 0 \\ \iff & \int_{-\infty}^{0} \delta(v)\,dv > 0 \\ \iff & \int_{-\infty}^{v^*} \delta(v)\,dv > -\int_{v^*}^{0} \delta(v)\,dv \\ \iff & \int_{-\infty}^{v^*} 1 - \Phi(\lambda(v))\,dv > \int_{v^*}^{0} \Phi(\lambda(v))\,dv \\ \iff & \int_{-\infty}^{v^*} (1 - \Phi(\lambda(x)))f_1(x)\,dx > \int_{v^*}^{0} \Phi(\lambda(x))f_1(x)\,dx \\ \impliedby & \int_{2v^*}^{v^*} (1 - \Phi(\lambda(x)))\,f_1(x)\,dx > \int_{v^*}^{0} \Phi(\lambda(x))\,f_1(x)\,dx \\ \impliedby & (1 - \Phi(\lambda(v^* - x)))\,f_1(v^* - x) > \Phi(\lambda(v^* + x))\,f_1(v^* + x) \qquad \forall x \in (0, -v^*) \\ \iff & f_1(v^* - x) > f_1(v^* + x) \qquad \forall x \in (0, -v^*) \checkmark \end{aligned}$$

Going line-by-line, the first implication is because $\delta(v) = 0$ for $v \geq 0$, and the second by splitting the integral at v^*. The third substitutes in the expression for $\delta(v)$, and the fourth does a change of variables to integrate against a uniform density rather than the density of project 1, $f_1(\cdot)$. The fifth truncates the left-hand integral at $2v^*$, which because the function being integrated is strictly positive, is strictly more restrictive than allowing the integral to go all the way to $-\infty$. The sixth examines the function pointwise, pairing each value $v^* - x$ on the left-hand side with its reflection $v^* + x$ about v^* on the right-hand side. The seventh uses the fact that v^* is precisely the value of v_1 for which $\mathbb{E}[v_2|v_1 = v] = \lambda(v) = 0$, and the standard normal $\Phi(\cdot)$ is symmetric about 0; thus the mass above $\lambda(v^* - x)$ is precisely equal to the mass below $\lambda(v^* + x)$. Finally, because $\mu_1 < 0$ and $\mu_2 < 0$, it must be that $v^* > \mu_1$ because there must be a positive shock to project 1 for either project to be approved. Therefore, because $f_1(v)$ is strictly decreasing in $|v - \mu_1|$ and $|(v^* - x) - \mu_1| < |(v^* + x) - \mu_1|$ for any $x > 0$, the last inequality does indeed hold for any $x \in (0, -v^*)$. □

B.2 Proofs of Main Results

This section provides proofs of the two main results: optimal discovery of trial balloons (Theorem 1) and disfavored projects in general (Theorem 2).

Before presenting those proofs, we first characterize optimal discovery in the straightforward case with no favored projects:

Proof of Proposition 7.

Let $w_1 = 1/2$. Then the principal's utility can be written as

$$\begin{aligned}
&\mathbb{P}(\text{both approved}) + w_1\,\mathbb{P}(\text{only 1 approved}) + w_2\,\mathbb{P}(\text{only 2 approved}) \\
&\qquad = \mathbb{P}(\text{both}) + \frac{1}{2}\,\mathbb{P}(\text{only 1}) + \frac{1}{2}\,\mathbb{P}(\text{only 2}) \\
&\qquad = \mathbb{P}(\text{both}) + \frac{1}{2}\,\mathbb{P}(\text{only 1 or only 2}) \\
&\qquad = \mathbb{P}(\text{both}) + \frac{1}{2}\,\big(\mathbb{P}(\text{at least 1}) - \mathbb{P}(\text{both})\big) \\
&\qquad = \frac{1}{2}\,\mathbb{P}(\text{both}) + \frac{1}{2}\,\mathbb{P}(\text{at least 1}).
\end{aligned}$$

From Propositions 5 and 6, we can see that both expressions are increasing more quickly in σ_1 when discovering project 1 than when discovering project 2. Furthermore, for $\sigma_1 < \sigma_2$ we know both terms are greater for discovering project 2, and for $\sigma_1 > (\mu_1/\mu_2)\,\sigma_2$ we know both terms are greater for discovering project 1. Thus there must be a cutoff $\underline{\sigma} \in (\sigma_2, (\mu_1/\mu_2)\,\sigma_2)$ such that discovering project 1 is best if and only if $\sigma_1 > \underline{\sigma}$. □

We are now ready to prove the first main result, which provides limit conditions for optimality of trial balloons.

Proof of Theorem 1.

We begin by decomposing the principal's utility as

$$\begin{aligned}
&\mathbb{P}(\text{both approved}) + w_1\,\mathbb{P}(\text{only 1 approved}) + w_2\,\mathbb{P}(\text{only 2 approved}) \\
&\qquad = w_1\,\mathbb{P}(\text{only 1 or both}) + w_2\,\mathbb{P}(\text{only 2 or both}) \\
&\qquad = w_1\,\mathbb{P}(\text{both}) + w_2\,\mathbb{P}(\text{belief}(v_2) \geq 0).
\end{aligned}$$

By belief(v_2), we refer to the agent's belief about the expected value of v_2, which may either be the actual value of v_2 (if project 2 was discovered) or the conditional expectation $\mathbb{E}[v_2|v_1]$ (if project 1 was discovered). The last line comes from noting two facts. First, because $\mu_1 < 0 < \mu_2$, regardless of which project is discovered, if project 1 would be approved alone then surely the grand bundle will be approved. Second, because $\mu_2 > 0$, then if the grand bundle would be approved then surely project 2 would be approved alone.

Let X be a standard normal random variable. Using the decomposition of the principal's payoffs, the payoff for discovering only project 1 is

$$\begin{aligned}
\Pi^1 &= w_1\,\mathbb{P}\big(v_1 + \mathbb{E}[v_2|v_1] \geq 0\big) + w_2\,\mathbb{P}\big(\mathbb{E}[v_2|v_1] \geq 0\big) \\
&= w_1\,\mathbb{P}\left(X > \frac{(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right) + w_2\,\mathbb{P}\left(X > \frac{-c\mu}{\rho\sigma_2}\right) \\
&= 1 - w_1\,\Phi\left(\frac{(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right) - w_2\,\Phi\left(\frac{-c\mu}{\rho\sigma_2}\right),
\end{aligned}$$

whereas the payoff for discovering only project 2 is

$$\begin{aligned}\Pi^2 &= w_1\,\mathbb{P}\big(v_2 + \mathbb{E}[v_1|v_2] \geq 0\big) + w_2\,\mathbb{P}\big(v_2 \geq 0\big) \\ &= w_1\,\mathbb{P}\left(X > \frac{(1-c)\mu}{\sigma_2 + \rho\sigma_1}\right) + \frac{1}{2}\,\mathbb{P}\left(X > \frac{-c\mu}{\sigma_2}\right) \\ &= 1 - w_1\Phi\left(\frac{(1-c)\mu}{\sigma_2 + \rho\sigma_1}\right) - w_2\,\Phi\left(\frac{-c\mu}{\sigma_2}\right).\end{aligned}$$

Because $\rho \in (0,1)$, it is always the case that $-c\mu/\sigma_2 > -c\mu/(\rho\sigma_2)$, and thus whenever $\sigma_1 \geq \sigma_2$, it is clear that $\Pi^1 \geq \Pi^2$.

We now compare discovering project 1 only to discovering neither:

$$\begin{aligned} w_2 &< 1 - w_1\,\Phi\left(\frac{(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right) - w_2\,\Phi\left(\frac{-c\mu}{\rho\sigma_2}\right) \\ 1 - w_1 &< 1 - w_1\,\Phi\left(\frac{(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right) - (1-w_1)\,\Phi\left(\frac{-c\mu}{\rho\sigma_2}\right) \\ (1-w_1)\,\Phi\left(\frac{-c\mu}{\rho\sigma_2}\right) &< w_1 - w_1\,\Phi\left(\frac{(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right) \\ \frac{1-w_1}{w_1}\,\Phi\left(\frac{-c\mu}{\rho\sigma_2}\right) &< 1 - \Phi\left(\frac{(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right) = \Phi\left(\frac{-(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right),\end{aligned}$$

where we use symmetry of the normal distribution to get the last equality. We observe that the left-hand side is decreasing in c whereas the right-hand side is increasing in c; furthermore, the expression is continuous. Therefore we have one of three cases:

1. The left-hand side is weakly greater than the right-hand side at $c = 1$ (and therefore for all $c \in [0,1)$ as well). This case occurs if and only if

$$\Phi\left(\frac{-\mu}{\rho\sigma_1}\right) > \frac{w_1}{2\,(1-w_1)}.$$

   Then, for fixed μ, discovering project 1 is worse than discovering nothing for all $c \in [0,1)$.

2. The left-hand side is weakly less than the right-hand side at $c = 0$ (and therefore for all $c \in [0,1)$ as well). This case occurs if and only if

$$\frac{1-w_1}{2w_1} < \Phi\left(\frac{-\mu}{\sigma_1 + \rho\sigma_2}\right).$$

   Then, for fixed μ, discovering project 1 is better than discovering nothing for all $c \in [0,1)$.

3. There is a single intersection at $\underline{c} \in (0,1)$; discovering project 1 is better than discovering nothing if and only if $c > \underline{c}$.

We start with the knife-edge case $w_1 = 1/2$, where the inequality reduces to

$$\Phi\left(\frac{-c\mu}{\rho\sigma_2}\right) < \Phi\left(\frac{-(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right),$$

which holds whenever

$$\frac{-(1-c)\mu}{\sigma_1+\rho\sigma_2} > \frac{-c\mu}{\rho\sigma_2} \iff c > \frac{\rho\sigma_2}{\sigma_1+2\rho\sigma_2} := c^*.$$

Thus the cutoff is interior and constant for all μ.

Now we apply the Implicit Function Theorem to the function

$$H(\mu, c(\mu)) = H(\mu, c) = \frac{1-w_1}{w_1}\,\Phi\!\left(\frac{-c\mu}{\rho\sigma_2}\right) - \Phi\!\left(\frac{-(1-c)\mu}{\sigma_1+\rho\sigma_2}\right)$$

in order to compute the derivative $c'(\mu)$ at the value $\underline{c}$ which produces equality, noting that we are guaranteed continuity of this function because H is continuous. The total derivative is

$$\left(\frac{1-w_1}{w_1}\,\frac{-c}{\rho\sigma_2}\,\phi\!\left(\frac{-c\mu}{\rho\sigma_2}\right) - \frac{-(1-c)}{\sigma_1+\rho\sigma_2}\,\phi\!\left(\frac{-(1-c)\mu}{\sigma_1+\rho\sigma_2}\right)\right) d\mu$$
$$+\left(\frac{1-w_1}{w_1}\,\frac{-\mu}{\rho\sigma_2}\,\phi\!\left(\frac{-c\mu}{\rho\sigma_2}\right) - \frac{\mu}{\sigma_1+\rho\sigma_2}\phi\!\left(\frac{-(1-c)\mu}{\sigma_1+\rho\sigma_2}\right)\right) dc,$$

and equating to 0 gives

$$\frac{dc}{d\mu} = \frac{-\frac{(1-w_1)\,c}{w_1\rho\sigma_2}\,\phi\left(\frac{-c\mu}{\rho\sigma_2}\right) + \frac{(1-c)}{\sigma_1+\rho\sigma_2}\,\phi\left(\frac{-(1-c)\mu}{\sigma_1+\rho\sigma_2}\right)}{\frac{(1-w_1)\mu}{w_1\rho\sigma_2}\,\phi\left(\frac{-c\mu}{\rho\sigma_2}\right) + \frac{\mu}{\sigma_1+\rho\sigma_2}\phi\left(\frac{-(1-c)\mu}{\sigma_1+\rho\sigma_2}\right)}.$$

The denominator is clearly positive, so we focus on the sign of the numerator, which is positive when

$$\frac{1-w_1}{w_1}\,\frac{c}{1-c}\,\frac{\sigma_1+\rho\sigma_2}{\rho\sigma_2}\,\phi\!\left(\frac{-c\mu}{\rho\sigma_2}\right) < \phi\!\left(\frac{-(1-c)\mu}{\sigma_1+\rho\sigma_2}\right)$$
$$\iff \frac{1-w_1}{w_1}\,\frac{c}{1-c}\,\frac{\sigma_1+\rho\sigma_2}{\rho\sigma_2} < \exp\left(-\left(\left(\frac{-(1-c)\mu}{\sigma_1+\rho\sigma_2}\right)^2 - \left(\frac{-c\mu}{\rho\sigma_2}\right)^2\right)\Big/ 2\right).$$

We first check the sign of the exponent, which is negative when

$$\frac{-(1-c)\mu}{\sigma_1+\rho\sigma_2} > \frac{-c\mu}{\rho\sigma_2} \iff c > \frac{\rho\sigma_2}{\sigma_1+2\rho\sigma_2} = c^*.$$

Note that for $w_1 < 1/2$, discovering policy 1 is always worse for the principal than if $w_1 = 1/2$, regardless of other parameters—reducing w_1 makes spoiling relatively more important. Thus in this case the cutoff $\underline{c}$ is bounded above $(\rho\sigma_2)/(\sigma_1+2\rho\sigma_2)$, so the exponent is negative at all relevant values of c. Furthermore, because $c/(1-c)$ is increasing in c, we can use the lower bound $\underline{c} \geq c^*$ to ensure the left-hand side is bounded above

$$\frac{1-w_1}{w_1}\,\frac{c^*}{1-c^*}\,\frac{\sigma_1+\rho\sigma_2}{\rho\sigma_2} = \frac{1-w_1}{w_1}.$$

Because $w_1 < 1/2$, the left-hand side is bounded strictly above 1. Thus because the right-hand side is bounded weakly below 1, the condition fails for all μ and the cutoff is decreasing.

If instead $w_1 > 1/2$, then by similar logic to the previous case, the exponent will be positive, so the right-hand side is bounded above 1; the left-hand side can be upper-bounded below 1 by using the upper bound $\underline{c} = c^*$ and simplifying; and we conclude that the cutoff is increasing.

We now finish our characterization. If $(1 - w_1)/(2w_1) < 1/2 \iff w_1 > 1/2$, then case (2) occurs for small enough μ, whereas case (1) never occurs because $w_1/(2 - 2w_1) > 1/2$. Thus for large w_1, there is $\bar{\mu} \in \mathbb{R}_+$ and an increasing cutoff $\underline{c}(\mu)$ such that $\underline{c}(\bar{\mu}) = 0$ and $\lim_{\mu\to\infty} \underline{c}(\mu) \leq (\rho\sigma_2)/(\sigma_1 + 2\rho\sigma_2)$. Discovering project 1 is best if and only if $c > \underline{c}(\mu)$.

If instead $w_1 < 1/2$ then case (1) attains for small μ and case (2) never attains, even in the limit. Then there is $\underline{\mu} \in \mathbb{R}_+$ and a decreasing cutoff $c^*(\mu)$ such that $\underline{c}(\underline{\mu}) = 1$ and $\lim_{\mu\to\infty} \underline{c}(\mu) \geq (\rho\sigma_2)/(\sigma_1 + 2\rho\sigma_2)$. Discovering project 1 is best if and only if $c > \underline{c}(\mu)$.

Finally, we know that in fact both limits must be precisely equal to c^*. In the limit, because the normal cdf has exponential tails, so the ratio

$$\Phi\left(\frac{-c\mu}{\rho\sigma_2}\right) / \Phi\left(\frac{-(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right)$$

is either 0 if $c > c^*$ and the numerator dominates, ∞ if $c < c^*$ and the denominator dominates, or 1 if $c = c^*$. Thus in order for this ratio to match the value $(1 - w_1/w_1)$ as μ grows large, the cutoff $\underline{c}$ must converge to c^* (and in fact must do so at an exponential rate). Because increasing w_1 (regardless of whether it is below or above 1/2) for fixed μ weakly decreases $\bar{c}$, it must be that the rate of convergence of $\underline{c}(\mu)$ to c^* is increasing in $|w_1 - 1/2|$. □

Next, we provide analogous limit conditions when the disfavored project is not a trial balloon.

Proof of Theorem 2.

For this theorem, we begin by proving two auxiliary results. The first is a corollary of Theorem 1 which applies that approach to compare discovering project 2 to no discovery:

Corollary 4. *Let $\mu_1 = -\mu < 0 \leq c\mu = \mu_2$ for $c \in [0, 1)$.*
For any μ, there is a cutoff $\underline{c}(\mu)$ such that it is better to discover project 2 than to discover neither project if and only if $c > \underline{c}(\mu)$.
The function $\underline{c}(\mu)$ is continuous and has the following properties:

1. *If $w_1 = 1/2$, then*
$$\underline{c}(\mu) = c^{**} := \frac{\sigma_2}{\rho\sigma_1 + 2\sigma_2} \qquad \text{for all } \mu.$$

2. *If $w_1 > 1/2$, there is $\mu^* \in \mathbb{R}_+$ such that $\underline{c}(\mu) = 0$ for all $\mu \in [0, \mu^*]$. For $\mu > \mu^*$, the cutoff $\underline{c}(\mu)$ is strictly increasing with $\lim_{\mu\to\infty} \underline{c}(\mu) = c^{**}$.*

3. *If $w_1 < 1/2$, there is $\mu^* \in \mathbb{R}_+$ such that $\underline{c}(\mu) = 1$ for all $\mu \in [0, \mu^*]$. For $\mu > \mu^*$, the cutoff $\underline{c}(\mu)$ is strictly decreasing with $\lim_{\mu\to\infty} \underline{c}(\mu) = c^{**}$.*

Proof. To compare discovery of project 2 to no discovery, we emulate the proof of Theorem 1. We omit the algebraic details, which are exactly as in that proof, and focus on the key conclusions.

The cutoff in the knife-edge case $w_1 = 1/2$ is given by

$$\frac{-(1-c)\mu}{\rho\sigma_1 + \sigma_2} > \frac{-c\mu}{\sigma_2} \iff c > \frac{\sigma_2}{\rho\sigma_1 + 2\sigma_2} := c^{**}.$$

Once again decreasing w_1 harms the principal (relative to no discovery) and increasing w_1 helps her, so the cutoff for $w_1 < 1/2$ is increasing in μ and converges to c^{**}, whereas the cutoff for w_2 is decreasing in μ and converges to c^{**}. □

The second result is a characterization of whether discovering project 1 or discovering project 2 dominates in the limit:

Lemma 14. *Let $\mu_1 = -\mu < 0 \leq c\mu = \mu_2$, for $c \in [0, 1)$ and let $\sigma_1 < \sigma_2$, so that project 1 is disfavored but not a trial balloon. Then,*

1. *For any $c \in (c^*, c^{**})$ there exists $\underline{\mu}$ such that discovering project 1 is better than discovering project 2 for all $\mu > \underline{\mu}$.*

2. *For any $c \in (c^{**}, 1)$ there exists $\bar{\mu}$ such that discovering project 2 is better than discovering project 1 for all $\mu > \bar{\mu}$.*

Proof. We make use of the expressions Π^1 and Π^2 from the proof of Theorem 1.

When $c \in (c^*, c^{**})$ we know from Theorem 1 that there exists μ' such that discovering project 1 is preferred to no discovery for all $\mu > \mu'$, and from Corollary 4 that there exists μ'' such that no discovery is preferred to discovering project 2 for all $\mu > \mu''$. Thus there exists $\bar{\mu} = \max\{\mu', \mu''\}$ such that discovering project 1 is preferred to discovering project 2. For $c < c^*$, we know no discovery will dominate so there is no need to establish a limit result.

When $c > c^{**}$, we know that discovering either project is eventually preferred to no discovery. However, because of the exponential nature of the normal distribution, we also know that the improvement probability from discovering project 2 dominates[1] not only the spoiling probability from discovering either project, but also the improvement probability from discovering project 1. Thus for μ large enough, discovering project 2 is better than discovering project 1. □

We now combine these two intermediate results to state the conclusion of the theorem. To make precise the cutoffs described in the text, we divide the result into the cases $w_1 = 1/2$, $w_1 < 1/2$, and $w_1 > 1/2$. We also describe a fourth region of very low μ for the latter two cases.

First, when $w_1 = 1/2$, the cutoffs are constant at c^* and c^{**}. Thus the low-μ case described in the main text vanishes. In the high-μ case, there is no discovery if $c < c^*$, discovery of project 1 if $c \in (c^*, c^{**})$, and discovery of an unknown project if $c > c^{**}$. In the limit case, there is no discovery for $c < c^*$, discovery of project 1 for $c \in (c^*, c^{**})$, and discovery of project 2 for $c > c^{**}$.

When $w_1 < 1/2$,

1. Very low μ: both cutoffs are 1, so there is no discovery for all c.

[1] By dominance we mean that the likelihood ratio between the improvement probability and the other specified probability becomes arbitrarily large.

2. Low μ: This region is lower-bounded by $\inf_\mu \left(\min \{c_1(\mu), c_2(\mu)\} < 1 \right)$. There is discovery for high enough c but it is unknown which project is discovered at each c.

3. High μ: This region is lower-bounded by the value of μ where $c_1(\mu) = c^{**}$. Because $c_1(\mu)$ is decreasing, there is no discovery for $c < c_1(\mu)$, discovery of project 1 for $c \in (c_1(\mu), c_2(\mu))$, and discovery of an unknown project for $c > c_2(\mu)$.

4. Limit μ: there is no discovery for $c < c^*$, discovery of project 1 for $c \in (c^*, c^{**})$, and discovery of project 2 for $c > c^{**}$.

Finally, when $w_1 > 1/2$,

1. Very low μ: both cutoffs are 0, so there is discovery of an unknown project for all c.

2. Low μ: This region is lower-bounded by $\inf_\mu \left(\max \{c_1(\mu), c_2(\mu)\} > 0 \right)$. There is discovery for high enough c but it is unknown which project is discovered at each c.

3. High μ: This region is lower-bounded by the value of μ where $c_2(\mu) = c^*$. Because $c_2(\mu)$ is increasing, there is no discovery for $c < c_1(\mu)$, discovery of project 1 for $c \in (c_1(\mu), c_2(\mu))$, and discovery of an unknown project for $c > c_2(\mu)$.

4. Limit μ: there is no discovery for $c < c^*$, discovery of project 1 for $c \in (c^*, c^{**})$, and discovery of project 2 for $c > c^{**}$.

In fact, in the high-μ cases, we can ensure that for some neighborhood of $c_2(\mu)$, discovery of project 1 is still optimal. This conclusion follows because when $c = c_2(\mu)$, the principal is precisely indifferent between discovering project 2 and no discovery. However, the principal strictly prefers discovering project 1 to no discovery. Because all utilities are continuous, it must be that for some ε-neighborhood above $c = c_2(\mu)$, discovering project 1 remains optimal. □

We now present a new result which establishes that the dominance region for project 1 contracts in the large-μ limit if $w_1 < 1/2$ and expands if $w_1 > 1/2$.

Lemma 15 (Limit Behavior of the Dominance Region for Project 1)**.** *Let $\mu_1 = -\mu < 0 \leq c\mu = \mu_2$ for $c \in [0, 1)$ and let $\sigma_1 < \sigma_2$. Then the following are true:*

- *There is $\underline{\mu}$ large enough such that, for all $\mu > \underline{\mu}$, the dominance region where project 1 is discovered differs by at most measure $\varepsilon > 0$ from $(c_1(\mu), c_2(\mu))$.*

- *For all $\mu > \underline{\mu}$, if $w_1 < 1/2$ then the measure of that dominance region is strictly decreasing in μ.*

- *For all $\mu > \underline{\mu}$, if $w_1 > 1/2$ then the measure of that dominance region is strictly increasing in μ.*

Proof of Lemma 15. Thanks to Lemma 14 we know that once $\mu > \underline{\mu}$, for any $c > c_2(\mu) \geq c^{**}$, project 2's greater improvement probability is the most important force driving disclosure; that is, it is first-order whereas all other effects are second-order. Thus we have two conclusions. First, the mass of $c > c_2(\mu)$ in the dominance region of project 1 is vanishing as μ gets large,

because only the improvement effect of project 2 will dominate the payoff of no discovery. This result is the first bullet point in the lemma. Second, it cannot be that some $c > c_2(\mu)$ is in the dominance region of project 1 for for $\mu > \underline{\mu}$ but not for $\mu' > \mu$, because the benefits of discovering project 1 will vanish faster than those of discovering project 2; the vanishing established in the first bullet point is therefore monotonic. We thus restrict attention to the behavior of the interval $(c_1(\mu), c_2(\mu))$.

To verify that the range of c where project 1 is discovered narrows with increasing μ when $w_1 < 1/2$, note that there are two effects: values $c < c_1(\mu)$ are added as $c_1(\mu)$ decreases and values $c > c_2(\mu)$ are removed as $c_2(\mu)$ decreases. In the limit, the rate of decrease of each is driven by the vanishing spoiling effect, so we can write

$$c_1'(\mu) \approx \frac{d}{d\mu}\Phi\left(\frac{-c\mu}{\rho\sigma_2}\right) = -\frac{c}{\rho\sigma_2}\,\phi\left(\frac{-c\mu}{\rho\sigma_2}\right) \qquad \text{and} \qquad c_2'(\mu) \approx \frac{d}{d\mu}\Phi\left(\frac{-c\mu}{\sigma_2}\right) = -\frac{c}{\sigma_2}\,\phi\left(\frac{-c\mu}{\sigma_2}\right).$$

In the limit the difference in the $\phi(\cdot)$ terms overcomes the difference in coefficients, and so $c_1(\mu)$ decreases slower than $c_2(\mu)$. Thus more values are removed from the range $(c_1(\mu), c_2(\mu))$ than added.

When $w_1 > 1/2$, values $c < c_2(\mu))$ are added as $c_2(\mu)$ increases and values $c < c_1(\mu)$ are removed as $c_1(\mu)$ increases. Both increases are driven by vanishing improvement, so we have

$$\begin{aligned} c_1'(\mu) &\approx \frac{d}{d\mu}\Phi\left(\frac{-(1-c)\mu}{\sigma_1+\rho\sigma_2}\right) = -\frac{(1-c)}{\sigma_1+\rho\sigma_2}\,\phi\left(\frac{-(1-c)\mu}{\sigma_1+\rho\sigma_2}\right) \qquad \text{and} \\ c_2'(\mu) &\approx \frac{d}{d\mu}\Phi\left(\frac{-(1-c)\mu}{\rho\sigma_1+\sigma_2}\right) = -\frac{(1-c)}{\rho\sigma_1+\sigma_2}\,\phi\left(\frac{-(1-c)\mu}{\rho\sigma_1+\sigma_2}\right), \end{aligned}$$

and $c_1(\mu)$ increases slower in the limit so the overall range where project 1 is discovered increases, similar to the

These properties may not hold away from the limit because we can no longer approximate the discovery region of project 1 by the interval $(c_1(\mu), c_2(\mu))$, and even if we could the rate of change of $c_1(\mu)$ and $c_2(\mu)$ depend on both spoiling and improvement rather than only one of the two effects. □

B.3 Proofs of Extensions

This section provides proofs and some additional technical discussion of the main extensions: multi-project discovery, noisy discovery, and discovery with more than two projects.

Proof of Theorem 3.

If $\mu_1 < 0$ and $\mu_2 < 0$, we may decompose utility as in Proposition 7 to get

$$\begin{aligned} &\mathbb{P}(\text{both approved}) + w_1\,\mathbb{P}(\text{only 1 approved}) + w_2\,\mathbb{P}(\text{only 2 approved}) \\ &\qquad = \frac{1}{2}\,\mathbb{P}(\text{both}) + \frac{1}{2}\,\mathbb{P}(\text{at least 1}). \end{aligned}$$

Then, as in the proof of that result, we may combine Propositions 5 and 6 to get the desired conclusion.

Now let $\mu_1 = -\mu < 0 \leq c\mu = \mu_2$ for $c \in [0, 1)$. Assuming $\sigma_1 \geq \sigma_2$ allows us to apply Theorem 1 and ignore the possibility of discovering project 2. We break this proof into two

lemmas. The first establishes dominance regions in the interval $[0,1)$ of possible values of c for discovering both projects or discovering only project 1. The second compares discovering both projects to no discover.

Lemma 16. *Let $\mu_1 = -\mu < 0 \leq c\mu = \mu_2$ for $c \in [0,1)$. Then, if*

$$c < c_\ell := \frac{\rho\sigma_2}{\rho\sigma_2 + \sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}},$$

there exists $\bar{\mu}$ such that discovering project 1 is better than full discovery for all $\mu > \bar{\mu}$. If instead

$$c > c_h := \frac{\sigma_2}{\sigma_1 + (1+\rho)\,\sigma_2},$$

then there exists $\underline{\mu}$ such that full discovery is better than discovering project 1 for all $\mu > \underline{\mu}$.

Proof. Let X be a $N(0,1)$ random variable. Using the decomposition above, the payoff for discovering both projects is

$$\begin{aligned}\Pi^b &= \frac{1}{2}\,\mathbb{P}\left(X < \frac{(1-c)\mu}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}}\right) - \frac{1}{2}\big(1 - \mathbb{P}(v_1 < 0 \text{ and } v_2 < 0)\big) \\ &\leq 1 - \frac{1}{2}\,\Phi\left(\frac{(1-c)\mu}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}}\right) - \frac{1}{2}\,\Phi\left(\frac{-c\mu}{\sigma_2}\right)\Phi\left(\frac{\mu}{\sigma_1}\right),\end{aligned}$$

The inequality comes from the fact that because the $\rho > 0$, the probability that both v_1 and v_2 are less than some value is bounded above by the product of the probabilities.

Because we cannot write the spoiling probability using a univariate normal cdf, we cannot follow the the approach of Lemma 14. Instead, we directly compare the expression Π^1 from Theorem 1 to our upper bound on Π^b:

$$\begin{aligned}1 - \frac{1}{2}\,\Phi\left(\frac{(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right) - \frac{1}{2}\,\Phi\left(\frac{-c\mu}{\rho\sigma_2}\right) &> 1 - \frac{1}{2}\,\Phi\left(\frac{(1-c)\mu}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}}\right) \\ &\quad - \frac{1}{2}\,\Phi\left(\frac{\mu}{\sigma_1}\right)\Phi\left(\frac{-c\mu}{\sigma_2}\right) \\ \iff \Phi\left(\frac{(1-c)\mu}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}}\right) + \Phi\left(\frac{\mu}{\sigma_1}\right)\Phi\left(\frac{-c\mu}{\sigma_2}\right) &> \Phi\left(\frac{(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right) + \Phi\left(\frac{-c\mu}{\rho\sigma_2}\right) \\ \iff \Phi\left(\frac{-c\mu}{\sigma_2}\right) - \Phi\left(\frac{-c\mu}{\rho\sigma_2}\right) - \left(1 - \Phi\left(\frac{\mu}{\sigma_1}\right)\right)\Phi\left(\frac{-c\mu}{\sigma_2}\right) &> \Phi\left(\frac{(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right) \\ &\quad - \Phi\left(\frac{(1-c)\mu}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}}\right) \\ \iff \int_{-c\mu/(\rho\sigma_2)}^{-c\mu/\sigma_2} \phi(x)\,dx - \left(1 - \Phi\left(\frac{\mu}{\sigma_1}\right)\right)\Phi\left(\frac{-c\mu}{\sigma_2}\right) &> \int_{(1-c)\mu/\sqrt{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}}^{(1-c)\mu/(\sigma_1+\rho\sigma_2)} \phi(x)\,dx.\end{aligned}$$

The term $(1 - \Phi(\mu/\sigma_1))\,\Phi(-c\mu/\sigma_2)$ vanishes with order $(\mu\exp(\mu^2))^2$. However, the integrals vanish only with order $\mu\exp(\mu^2)$. Thus for μ sufficiently large we can directly compare the

integrals. We also use the symmetry of the normal distribution to switch from the integral over $(-c\mu/(\rho\sigma_2), -c\mu/\sigma_2)$ to the integral over $(c\mu/\sigma_2, c\mu/(\rho\sigma_2))$:

$$\begin{aligned}
&\int_{c\mu/\sigma_2}^{c\mu/(\rho\sigma_2)} \phi(x)\,dx > \int_{(1-c)\mu/\sqrt{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}}^{(1-c)\mu/(\sigma_1+\rho\sigma_2)} \phi(x)\,dx \\
\Longleftarrow\ &\phi\left(\frac{c\mu}{\rho\sigma_2}\right) c\,\mu\,\frac{1-\rho}{\rho\sigma_2} > \phi\left(\frac{(1-c)\mu}{\sqrt{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}}\right)(1-c)\,\mu\left(\frac{1}{\sigma_1+\rho\sigma_2} - \frac{1}{\sqrt{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}}\right) \\
\Longleftrightarrow\ &\phi\left(\frac{c\mu}{\rho\sigma_2}\right)\Big/ > \frac{1-c}{c}\,\frac{\rho\sigma_2}{1-\rho}\left(\frac{1}{\sigma_1+\rho\sigma_2} - \frac{1}{\sqrt{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}}\right)\phi\left(\frac{(1-c)\mu}{\sqrt{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}}\right) \\
&\qquad := k \\
&\Longleftrightarrow\ k < \exp\left(-\left(\left(\frac{c\mu}{\rho\sigma_2}\right)^2 - \left(\frac{(1-c)\mu}{\sqrt{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}}\right)^2\right)\Big/ 2\right) \\
\Longleftarrow\ &\text{for large } \mu \qquad \left(\frac{c\mu}{\rho\sigma_2}\right)^2 < \left(\frac{(1-c)\mu}{\sqrt{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}}\right)^2 \\
&\Longleftrightarrow\ c < \frac{\rho\sigma_2}{\rho\sigma_2+\sqrt{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}} =: c_\ell.
\end{aligned}$$

We can now re-introduce the term $(1-\Phi(\mu/\sigma_1))\,\Phi(-c\mu/\sigma_2)$. Our limit result states that for μ large enough, the inequality in the first line holds. Beyond this point, the difference between the integrals vanishes with the same order as each integral on its own; that is, with order $\mu\exp(\mu^2)$. Thus the term $(1-\Phi(\mu/\sigma_1))\,\Phi(-c\mu/\sigma_2)$ will be exponentially smaller than the difference between the integrals and does not alter the inequality.

Now we perform the reverse bounding exercise to show that discovering both projects is better than discovering project 1 for high c. In place of Π^b we use a lower bound on the utility from discovering both, obtained by subtracting $\mathbb{P}(v_2 < 0)$ rather than $\mathbb{P}(v_1 < 0 \text{ and } v_2 < 0)$.

Using $\mathbb{P}(v_1 < 0)$ would result in a looser bound.

$$
\begin{aligned}
1 - \frac{1}{2}\Phi\left(\frac{(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right) - \frac{1}{2}\Phi\left(\frac{-c\mu}{\rho\sigma_2}\right) &< 1 - \frac{1}{2}\Phi\left(\frac{(1-c)\mu}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}}\right) - \frac{1}{2}\Phi\left(\frac{-c\mu}{\sigma_2}\right) \\
\iff \Phi\left(\frac{(1-c)\mu}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}}\right) + \Phi\left(\frac{-c\mu}{\sigma_2}\right) &< \Phi\left(\frac{(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right) + \Phi\left(\frac{-c\mu}{\rho\sigma_2}\right) \\
\iff \Phi\left(\frac{-c\mu}{\sigma_2}\right) - \Phi\left(\frac{-c\mu}{\rho\sigma_2}\right) &< \Phi\left(\frac{(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right) - \Phi\left(\frac{(1-c)\mu}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}}\right) \\
\iff \int_{-c\mu/(\rho\sigma_2)}^{-c\mu/\sigma_2} \phi(x)\,dx &< \int_{(1-c)\mu/\sqrt{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}}^{(1-c)\mu/(\sigma_1+\rho\sigma_2)} \phi(x)\,dx \\
\iff \int_{c\mu/\sigma_2}^{c\mu/(\rho\sigma_2)} \phi(x)\,dx &< \int_{(1-c)\mu/\sqrt{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}}^{(1-c)\mu/(\sigma_1+\rho\sigma_2)} \phi(x)\,dx \\
\impliedby \frac{1}{k} &< \phi\left(\frac{(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right) \Big/ \phi\left(\frac{c\mu}{\sigma_2}\right) \\
\iff \frac{1}{k} &< \exp\left(-\left(\left(\frac{(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right)^2 - \left(\frac{c\mu}{\sigma_2}\right)^2\right)\Big/ 2\right) \\
\impliedby \text{for large } \mu \qquad \left(\frac{(1-c)\mu}{\sigma_1 + \rho\sigma_2}\right)^2 &< \left(\frac{c\mu}{\sigma_2}\right)^2 \\
\iff c_h := \frac{\sigma_2}{\sigma_1 + (1+\rho)\,\sigma_2} &< c.
\end{aligned}
$$

Note the similarity of this expression to c^* and c^{**} in Theorem 2. The difference comes from comparing the improvement probability associated with discovering project 1 to the spoiling probability associated with discovering project 2. Thus this bound lies above c^* (because spoiling is stronger than when discovering project 1) but below c^{**} (because improvement is weaker than when discovering project 2). □

The second lemma compares no discovery to discovering both projects:

Lemma 17. *Let $\mu_1 = -\mu < 0 \leq c\mu = \mu_2$ for $c \in [0, 1)$. There is a cutoff c_{no} such that*

- *For any $c \in [0, c_{no})$, there exists $\bar{\mu}(c)$ such that full discovery is better than no discovery if and only if $\mu \in [0, \bar{\mu}(c))$.*
- *For any $c > c_{no}$, full discovery is better than no discovery for all μ.*

Proof. Studying the contour sets of the normal distribution shows that for any $c < 1/2$, discovering neither is eventually better than discovering both, and once that occurs, that dominance is monotonic. For $\rho \in (0, 1)$, the cutoff obtained for discovering project 1 to dominate discovering both projects is below $1/2$ so this argument always applies.

Formally, we can plot the contour sets of the standard normal in μ_1-μ_2 space as ellipses slanted along the line $y = x$ and centered at the origin. Discovering both projects can be seen as choosing a contour set, then drawing uniformly from the contour set. On each contour set, we can draw three regions:

1. Above the line $y = -x$, where the sum of project values is positive.

2. In the third quadrant, where both project values are negative.

3. The remaining region (two symmetric triangles in the first and fourth quadrants) where exactly one project value is positive.

For the standard normal, we can see that this decomposition yields a positive expected value for each contour set.

The contour sets of a normal with different means can be derived by shifting the center appropriately. As we shift the center to the left along the x-axis, meaning that project 1's mean becomes more negative, regions (2) and (3) expand to cover region (1) without increasing the relative share of region (3) compared to region (2). Thus expected payoff decreases. If we instead shift the center up along the y-axis, meaning that project 2's mean becomes more positive, regions (1) and (3) expand to cover region (2) without increasing the relative share of region (3) compared to region (1). Thus expected payoff increases. Therefore, as we move along some curve $c_{\text{no}} = c(\mu)$ in the second quadrant, regions (1) and (2) become and remain equal in size for all ellipses not containing the origin. For such contour sets, the expected payoff is 1/2; however, because the distribution has full support, some contour sets will cover the origin. The payoff in those contour sets is strictly larger, because in each of them region (1) is larger than region (2). Thus the payoff from full discovery for any project with mean lying on the curve c_{no} exceeds 1/2, the payoff from no discovery. For $c > c_{\text{no}}$, the payoff from revealing both increases, so full discovery continues to dominate. For $c < c_{\text{no}}$, the region (2) is larger than the region (1) for ellipses not containing the origin. For μ large, mass is concentrated in these ellipses; because the normal distribution has vanishing tails, we can eventually ignore the mass from ellipses containing the origin (i.e., for any $\varepsilon > 0$ there is μ large enough that the gains from full discovery captured by the origin-containing contour sets are below ε). By taking μ large enough that those ε-gains are bounded below the losses from the contour sets not containing the origin, we ensure that full discovery is worse than no discovery. Therefore for any $c < c_{\text{no}}$, there is μ large enough that no discovery is better than no discovery. The payoffs from full discovery are monotonically declining in μ, so in fact for $c < c_{\text{no}}$ there is a cutoff $\bar{mu}(c)$ such that full discovery is preferred if and only if $\mu \in [0, \bar{\mu})$. □

We then combine the two lemmas to state the theorem. Note that we do not bound c_{no} in relation to c_ℓ and c_h, or indeed provide a closed-form solution, leading to the presentation in the text. □

Proof of Proposition 8.

We begin by computing the distribution of the agent's posterior mean belief about v_i, which we call q_i, after observing the signal s_i. Because s_i is normally distributed, as is the prior over v_i, we can use standard properties of the normal distribution to get

$$q_i \sim N\left(\mu_i, \frac{\sigma_i^4}{\tau_i^2 + \sigma_i^2}\right).$$

Intuitively, as the signal s_i becomes noisier, the agent pays less attention to it, so beliefs do not stray far from the prior mean μ_i. In the limiting case $\tau_i^2 = 0$, the distribution is the same

as if the agent observed public discovery of v_i (as in Section 2.4). In the other limiting case $\tau_i^2 \to \infty$, the signal is completely uninformative and the agent always places full weight on the prior mean μ_i.

Although this distribution describes the possible posterior mean beliefs after observing s_i, note that the agent's actual posterior about v_i will have distribution

$$v_i \sim N(q_i, \sigma_i^2 + \tau_i^2).$$

Thus we get the distribution

$$v_j|s_i \sim N\left(\mu_j + \rho\frac{\sigma_j}{\sqrt{\sigma_i^2+\tau_i^2}}(s_i - \mu_i), (1-\rho)^2\sigma_j^2\right)$$

for the agent's posterior belief about v_j.

Thus we can replicate all the results of Section 2.4 using the distribution

$$N\left(\mu_i, \frac{\sigma_i^4}{\tau_i^2+\sigma_i^2}\right)$$

for the project i which the principal chooses to discover and

$$N\left(\mu_j + \rho\frac{\sigma_j}{\sqrt{\sigma_i^2+\tau_i^2}}(s_i - \mu_i), (1-\rho)^2\sigma_j^2\right)$$

for the posterior belief about the project j conditional on discovery of project i. A noisy signal weakens both the spoiling and improvement effects for project i by reducing the amount by which discovery shifts the agent's beliefs. □

Proof of Proposition 9.

We begin by characterizing the principal's utility in a general N-project setting, with no restrictions on project means or variances. Extending the computation in the proof of Proposition 5, we see that the probability of obtaining a bundle containing projects $j \in S$ for some subset of projects S, upon revealing project i, is

$$\begin{aligned}\mathbb{P}\left(v_i + \sum_{j\in S}\mathbb{E}[v_j|v_i] \geq 0\right) &= \mathbb{P}\left(v_i \geq \frac{\rho\mu_i\sum_{j\in S}(\sigma_j/\sigma_i) - \sum_{j\in S}\mu_j}{1+\rho\sum_{j\in S}(\sigma_j/\sigma_i)}\right)\\ &= \Phi\left(\frac{\mu_i + \sum_{j\in S}\mu_j}{\sigma_i + \rho\sum_{j\in S}\sigma_j}\right)\end{aligned}$$

Similarly, extending the decomposition of utility from the proof of Theorem 1, we get that the principal's utility is given by

$$\sum_{i\in N} w_i\,\mathbb{P}(\text{some bundle containing } i).$$

Define S_j^i as

$$S_j^i := \left(\arg\max_{S\subseteq N}\mathbb{P}(\text{bundle containing } j \text{ approved} \mid i \text{ discovered})\right) \setminus \{i,j\},$$

that is, the most likely bundle containing j conditional on project i being discovered, excluding project i and j themselves. To further parallel Theorem 1, let $\mu_i = -\mu < 0$ and let $\mu_j = c_j \mu$ for constants $c_j \in \mathbb{R}$. Then we can write the principal's utility from discovering project i as

$$w_i \, \Phi\left(\frac{-(1 - \sum_{j \in S_i^i} c_j)\,\mu}{\sigma_i + \rho \sum_{j \in S_j^i} \sigma_j}\right) + \sum_{j \text{ s.t. } c_j < 0} w_j \, \Phi\left(\frac{-(1 - c_j - \sum_{j \in S_j^i} c_j)\,\mu}{\sigma_i + \rho\sigma_j + \rho \sum_{j \in S_j^i} \sigma_j}\right) - \sum_{j \text{ s.t. } c_j \geq 0} w_j \, \Phi\left(\frac{-c_j \mu}{\rho \sigma_j}\right).$$

An approach in the style of Corollary 4 would allow us to compute the utility from discovering a project j with $c_j \geq 0$ as well.

From this expression, we can see that the approach to comparing discovery of project i with no discovery is largely preserved. Fixing all parameters except some c_j, the first two terms are weakly increasing in c_j whereas the third is weakly decreasing in c_j. Thus there is once again a well-defined continuous cutoff function $\underline{c}(\mu) \in \mathbb{R}$ such that the sum of the first two terms exceeds the third if and only if $c_j > \underline{c}_j(\mu)$. Unfortunately, we cannot analytically sign $\underline{c}_j'(\mu)$ to determine whether the range of viable c_j increases or decreases with higher precision; we expect that the result will be highly dependent on the remaining parameters of the model.

Now we impose the assumptions stated in the text to simplify the expression above:

$$w_i \, \Phi\left(\frac{-(1 - \sum_{j \in S_i^i} c_j)\,\mu}{\sigma_i + \rho \sum_{j \in S_i^i} \sigma_j}\right) + \sum_{j \neq i} w_j \, \Phi\left(\frac{-c_j \mu}{\rho \sigma_j}\right).$$

If another project k is discovered, we have utility

$$w_i \, \Phi\left(\frac{-(1 - c_k - \sum_{j \in S_i^k} c_j)\,\mu}{\rho\sigma_i + \sigma_k + \rho \sum_{j \in S_i^k} \sigma_j}\right) + \Phi\left(\frac{-c_k \mu}{\sigma_k}\right) + \sum_{j \notin \{i,k\}} w_j \, \Phi\left(\frac{-c_j \mu}{\rho \sigma_j}\right).$$

Discovering project i is preferred whenever

$$w_i \, \Phi\left(\frac{-(1 - \sum_{j \in S_i^i} c_j)\,\mu}{\sigma_i + \rho \sum_{j \in S_i^i} \sigma_j}\right) + w_k \, \Phi\left(\frac{-c_k \mu}{\rho \sigma_j}\right) > w_i \, \Phi\left(\frac{-(1 - c_k - \sum_{j \in S_i^k} c_j)\,\mu}{\rho\sigma_i + \sigma_k + \rho \sum_{j \in S_i^k} \sigma_j}\right) + \Phi\left(\frac{-c_k \mu}{\sigma_k}\right).$$

Clearly the second term on the left-hand side is larger than the second term on the right-hand side, as in the proof of Theorem 1. If project i also satisfies $\sigma_i = \max_{j \in N} \sigma_j$, then we can show that the first term on the left-hand side is greater than the first term on the right-hand side, so discovering project i is always better than discovering any other project:

$$\frac{-(1 - \sum_{j \in S_i^i} c_j)\,\mu}{\sigma_i + \rho \sum_{j \in S_i^i} \sigma_j} \geq \frac{-(1 - c_k - \sum_{j \in S_i^k} c_j)\,\mu}{\sigma_i + \rho\sigma_k + \rho \sum_{j \in S_i^k} \sigma_j} \geq \frac{-(1 - c_k - \sum_{j \in S_i^k} c_j)\,\mu}{\rho\sigma_i + \sigma_k + \rho \sum_{j \in S_i^k} \sigma_j}.$$

The first inequality follows because S_i^i is by construction a better choice of bundle when discovering project i than $S_i^k \cup \{k\}$. The second follows from $\sigma_i \geq \sigma_k$. Thus discovering a trial balloon remains optimal in this setting.

Finally, we examine the content of Theorem 2 in this setting. If project i remains the only negative-mean project but $\sigma_i < \max_{j \in N} \sigma_j$, then we cannot in general sign the value

$$\delta = \left(\frac{-(1 - \sum_{j \in S_i^i} c_j)\,\mu}{\sigma_i + \rho \sum_{j \in S_i^i} \sigma_j}\right) - \left(\frac{-(1 - c_k - \sum_{j \in S_i^k} c_j)\,\mu}{\rho\sigma_i + \sigma_k + \rho \sum_{j \in S_i^k} \sigma_j}\right).$$

If $\delta > 0$ for all parameter values despite the assumption on σ_i, then Proposition 9 still holds. Otherwise, in the cases where $\delta < 0$ we have the same trade-off as in Theorem 2, where project i has a lower spoiling chance than any project k, but project k has a greater improvement chance. We can find limit cutoffs in the style of c^* and c^{**} in that theorem for each c_j with $j \in S_i^i \cup S_i^k \cup \{k\}$, and compare them if $c_j \in (S_i^i \cup \{k\}) \cap (S_i^k \cup \{k\})$ to obtain a version of Theorem 2 in this more general setting. However, it is harder to visualize these results in the style of Figure 2.2 because the space $\{c_j\} \times \mu$ is now N-dimensional.

If project i is no longer the only negative-mean project, we now have to deal with the potential for $S_j^i \neq S_j^k$ for all negative-mean projects j. That is, we now have an expression δ_j in the style of δ above comparing improvement effects across discovery rules for each negative-mean project. In general, we cannot sign these expressions without explicitly fixing parameters, and cannot even make a comparison of limit cutoffs using the techniques of Theorem 2 because there are multiple improvement effects to consider. □

Appendix C

Additional Material for "Strategic Opinion-Writing on Appellate Courts"

Tables and Figures from the Main Text

Table C.1: Summary Statistics for Opinion Text

Type	Total	Has Text	Known Author	> 500 Words	> 1000 Words	> 5000 Words
Lead/Majority	387,898	313,745	274,416	251,720	224,270	24,842
Dissenting	31,433	31,116	30,734	22,431	15,746	1,694
Concurring	21,361	16,005	16,005	8,499	5,329	560

My sample contains the universe of published cases from U.S. federal courts of appeals, 1890–2013. Of particular interest is the period 1970–2013, when ideology scores are available for all judges and partisan legal positions are relatively stable; this time period comprises approximately 70% of cases for which I have metadata.

Table C.2: Polarization Causes Total Dissents to Increase

	(1) Party	(2)	(3) JCS Scores	(4)	(5)	(6) DIME-CF Scores	(7)
	All Panels	All Panels	Party-Unan.	Bip.	All Panels	Party-Unan.	Bip.
bipartisan	0.232^{***}						
	(0.023)						
extremal dist.		0.610^{***}	0.096	0.381^{***}	0.223^{***}	0.021	0.065^{*}
		(0.049)	(0.150)	(0.084)	(0.025)	(0.062)	(0.035)
2nd-extremal dist.		0.088	−0.117	0.056	0.169^{***}	−0.106	0.129^{**}
		(0.086)	(0.271)	(0.103)	(0.055)	(0.159)	(0.063)
p for coeff. diff.	–	0.000	0.527	0.013	0.393	0.487	0.409
Control dissents	0.084	0.077	0.077	0.077	0.079	0.078	0.077
Score std. dev.	–	0.388	0.382	0.386	0.704	0.672	0.697
Observations	257, 712	150, 841	39, 095	89, 140	155, 129	40, 471	91, 595

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Results from conditional logit estimation of dissent rates using circuit × year fixed effects (robust standard errors clustered at the circuit × year level), following Equation (3.2) for column (1) and Equation (3.3) for columns (2)–(7). The latter group of columns is divided into three estimates using JCS scores to measure distances between judges—columns (2)–(4)—and three columns repeating the analysis with DIME-CF scores—columns (5)–(7). p-values test for equality between extremal distance and second-extremal distance coefficients using a Wald test. Control dissent rates for columns using scores are computed by averaging panels in the sample where the extremal distance to the median is < 0.25.

Table C.3: Polarization Causes Equally Increased Dissents Across Judges

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		JCS Scores				DIME-CF Scores			
Outcome	Variable	All Panels	All Panels	Party-Unan.	Bip.	All Panels	All Panels	Party-Unan.	Bip.
dissent by median judge	bipartisan	0.320***				0.310***			
		(0.035)				(0.035)			
	ext. dist.		0.542***	0.329	0.330***		0.229***	0.114	0.067
			(0.057)	(0.209)	(0.099)		(0.031)	(0.085)	(0.045)
	2nd-ext. dist.		0.171	−0.279	0.112		0.137**	−0.145	0.116
			(0.119)	(0.423)	(0.144)		(0.069)	(0.206)	(0.082)
dissent by 2nd-extremal judge	bipartisan	0.333***				0.322***			
		(0.035)				(0.035)			
	ext. dist.		0.503***	0.014	0.299***		0.182***	−0.094	0.057
			(0.057)	(0.213)	(0.100)		(0.032)	(0.095)	(0.046)
	2nd-ext. dist.		0.040	0.037	−0.029		0.212***	0.189	0.129
			(0.120)	(0.414)	(0.144)		(0.070)	(0.207)	(0.082)
dissent by extremal judge	bipartisan	0.311***				0.308***			
		(0.035)				(0.035)			
	ext. dist.		0.565***	−0.053	0.362***		0.178***	0.020	0.042
			(0.057)	(0.206)	(0.100)		(0.032)	(0.089)	(0.046)
	2nd-ext. dist.		0.025	−0.144	0.079		0.106	−0.340	0.100
			(0.118)	(0.408)	(0.140)		(0.070)	(0.209)	(0.082)
	bip./ext. = across roles	0.907	0.747	0.398	0.910	0.955	0.463	0.275	0.929
	2nd-ext. = across roles	–	0.644	0.870	0.777	–	0.557	0.203	0.971
	Control dissents	0.042	0.076	0.076	0.096	0.044	0.078	0.073	0.095
	Score std. dev.	–	0.388	0.351	0.397	–	0.704	0.666	0.709
	Observations	248, 693	150, 900	40, 271	89, 295	249, 043	155, 191	41, 530	91, 768

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Results from conditional multinomial logit estimation of dissent rates using circuit × year fixed effects (with robust standard errors clustered at the circuit × year level), following Equation (3.4) for column (1) and (5) and Equation (3.5) for columns (2)–(4) and (5)–(8). Columns (1)–(4) use JCS scores to determine median, second-extremal, and extremal judges and to measure distances to the median; columns (5)–(8) repeat that analysis using DIME-CF scores. p-values test for equality between coefficients across judge roles using a Wald test. Control dissent rates for columns using scores are computed as in Table C.2, by averaging panels in the sample where the extremal distance to the median is < 0.25.

Table C.4: Outliers Cause Minimal Differences in Judge-Level Dissents

	(1) Min. Party	(2) Is JCS Median	(3) Is DIME-CF Median	(4) Min. Law School	(5) Min. Gender	(6) Only Man	(7) Only Woman	(8) Youngest	(9) Median Age	(10) Oldest
judge char.	0.010	0.014	−0.018	0.017	0.126***	−0.037	0.144***	0.020	0.009	0.004
	(0.022)	(0.021)	(0.023)	(0.043)	(0.041)	(0.072)	(0.045)	(0.022)	(0.017)	(0.023)
Control dissents	0.031	0.036	0.036	0.031	0.038	0.033	0.032	0.032	0.033	0.033
Observations	520, 148	379, 597	388, 432	175, 973	138, 036	138, 036	138, 036	772, 230	772, 230	772, 230
# judges	1, 230	829	841	824	648	648	648	1,370	1,370	1,370

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Results from conditional logit estimation of dissent rates at the judge level using circuit × year and judge fixed effects (with robust standard errors clustered at the judge level), following Equation (3.6). All specifications include controls for Bloomberg topic, as well as number of Democrats, number of unique law schools, and number of women whenever those are not the characteristic of interest. The sample for each column includes only mixed-characteristic panels (e.g., in column (1), only bipartisan panels). The control dissent rate is computed at the judge level across included panels where the judge does not satisfy the criterion of interest (e.g., in column (1), over judges on bipartisan panels that are not in the partisan minority).

Table C.5: Majority Dissent Shares Are Nearly Proportional

	(1) Maj. Party	(2) Non-Med. JCS	(3) Non-Med. DIME-CF	(4) Maj. Law School	(5) Maj. Gender	(6) Maj. Man	(7) Maj. Woman	(8) Youngest	(9) Median Age	(10) Oldest
w/ controls	0.663	1.003	1.003	0.666**	0.631	0.621	0.682	1.017	1.018	0.983
	(0.062)	(0.120)	(0.119)	(0.114)	(0.083)	(0.086)	(0.178)	(0.127)	(0.111)	(0.129)
	$p = 0.394$	$p = 0.139$	$p = 0.172$	$p = 0.038$	$p = 0.450$	$p = 0.471$	$p = 0.361$	$p = 0.510$	$p = 0.493$	$p = 0.484$
w/o controls	0.663	1.004	1.001*	0.667**	0.637	0.630	0.657	1.011	1.018	0.985
	(0.062)	(0.120)	(0.118)	(0.110)	(0.080)	(0.083)	(0.163)	(0.124)	(0.108)	(0.127)
	$p = 0.374$	$p = 0.201$	$p = 0.057$	$p = 0.033$	$p = 0.482$	$p = 0.484$	$p = 0.206$	$p = 0.444$	$p = 0.503$	$p = 0.450$
Control dissents	0.083	0.064	0.065	0.066	0.090	0.090	0.132	0.093	0.094	0.093
Observations	261,364	164,593	164,878	76,333	290,447	281,977	6,857	263,097	287,655	260,019

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Ratios of majority characteristic dissents on bi-characteristic panels computed as described in Section 3.6.3. Coefficients used to compute the level of dissents on bipartisan panels are estimated using a conditional logit following Equation (3.2) or the analogous Equation (3.10), with circuit × year fixed effects and robust standard errors clustered at the circuit × year level. The row with controls always includes Bloomberg topic, as well as number of Democrats, number of unique law schools, and number of women whenever those are not the characteristic of interest. Standard errors are computed using 1,000 Bayesian bootstrap draws of the data. p-values are for a two-sided test of equality with the proportional value (2/3 for columns (1) and (4)–(7); 1 for the other columns) also conducted using 1,000 Bayesian bootstrap draws of the data.

Table C.6: Equally Increased Dissents Persist Across Legal Topics

	(1) Crim.	(2) Civ. Rights	(3) 1st Am.	(4) Due Process	(5) Labor Rel.	(6) Econ. & Reg.	(7) Misc.
median dis.	0.367^{***}	0.291^{***}	0.137	0.311^{***}	0.528^{***}	0.142^{*}	0.472^{***}
	(0.074)	(0.108)	(0.331)	(0.065)	(0.143)	(0.076)	(0.129)
2nd-ext. dis.	0.444^{***}	0.717^{***}	0.166	0.239^{***}	0.374^{***}	0.160^{**}	0.388^{***}
	(0.072)	(0.118)	(0.389)	(0.064)	(0.138)	(0.076)	(0.124)
ext. dis.	0.370^{***}	0.466^{***}	−0.056	0.339^{***}	0.155	0.146^{*}	0.429^{***}
	(0.072)	(0.116)	(0.385)	(0.065)	(0.130)	(0.075)	(0.127)
p for coeff. diff.	0.701	0.032	0.906	0.540	0.156	0.985	0.898
Control dissents	0.046	0.076	0.135	0.054	0.055	0.030	0.065
Observations	56, 825	14, 744	1, 675	61, 533	15, 766	82, 382	15, 479

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Results from conditional multinomial logit estimation of dissent rates using circuit × year fixed effects (with robust standard errors clustered at the circuit × year level), following Equation (3.4). Each column restricts the sample to one of the eight Bloomberg Law topics (dropping privacy cases, with a sample of < 500, because of co-linearity preventing estimation). Coefficients represent the increase in dissents by each judge role on bipartisan panels compared to that same judge role on party-unanimous panels, i.e., the same specification as column (1) of Table C.3 under the sample restriction. p-values test for equality between coefficients across judge roles using a Wald test.

Table C.7: Outliers Weakly Increase Total Dissents

	(1) Partisanship	(2) Shared Law School	(3) Any Law Schools	(4) Mixed Gender	(5) Min. Men	(6) Min. Women	(7) Age
bipartisan	0.232^{***}						
	(0.023)						
one diff. law school		0.056	−0.024				
		(0.064)	(0.021)				
all same law school			−0.079				
			(0.066)				
mixed gender				0.076^{***}			
				(0.026)			
one man					0.083		
					(0.165)		
one woman						0.067^{***}	
						(0.025)	
young outlier							0.041
							(0.028)
old outlier							0.040
							(0.030)
Control dissents	0.084	0.066	0.096	0.094	0.091	0.090	0.090
Observations	257, 712	70, 875	257, 712	257, 712	5, 454	251, 574	257, 712

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Results from conditional logit estimation of outlier effects on total dissents using circuit × year fixed effects (with robust standard errors clustered at the circuit × year level), following Equation (3.7). For columns (1), (3), (4), and (7) all panels are included. For column (2), all panels with two or more judges from the same law school are included. For column (5), all majority-women panels are included. For column (6), all majority-men panels are included. On panels with a "young outlier," the youngest judge is 10 years or more younger than the median judge, who is in turn less than 5 years younger than the oldest judge. Panels with an "old outlier" have the same phenomenon but with the age ordering reversed.

Table C.8: Outliers Cause Concentrated Increases in Dissents Across Judges

	(1) Three Unique Law Schools	(2) Any Age Outlier	(3) Young Outlier	(4) Old Outlier
second-extremal or youngest judge	0.022	0.024	−0.000	0.046
	(0.028)	(0.031)	(0.040)	(0.042)
median judge	0.046	0.009	0.016	−0.001
	(0.028)	(0.031)	(0.039)	(0.042)
extremal or oldest judge	0.025	0.072^{**}	0.076^{**}	0.049
	(0.028)	(0.030)	(0.038)	(0.042)
p, all equal	0.819	0.322	0.361	0.649
Control dissents	0.099	0.094	0.093	0.093
Observations	250, 595	259, 821	259, 821	259, 821

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Results from conditional multinomial logit estimation of dissent rates using circuit × year fixed effects (with robust standard errors clustered at the circuit × year level), following Equation (3.8) for column (1) and Equation (3.9) for columns (2)–(4). The sample for column (1) includes only panels with at least two unique law schools; coefficients represent the increase in dissents on panels with three unique law schools compared to two. In columns (2)–(4), coefficients represent the increase in dissents on panels where the specified age outlier is present. The reported p-values test for equality between coefficients within a column using a Wald test.

Table C.9: Partisan Polarization Slightly Increases Minority-Characteristic Dissents

	(1) Law School	(2) Gender (Pooled)	(3) Gender (Men)	(4) Gender (Women)	(5) Age
minority / youngest	0.343^{***}	0.253^{***}	1.096^{***}	0.141^{*}	0.345^{***}
	(0.083)	(0.073)	(0.245)	(0.078)	(0.038)
majority / age-median	0.250^{***}	0.285^{***}	0.258^{***}	0.426^{***}	0.312^{***}
	(0.059)	(0.056)	(0.062)	(0.141)	(0.039)
oldest	–	–	–	–	0.316^{***}
					(0.039)
p for coeff. diff.	0.372	0.740	0.001	0.079	0.806
Control dissents	0.080	0.095	0.095	0.095	0.083
Observations	65,368	48,610	48,610	48,610	259,821

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Results from conditional multinomial logit estimation of dissent rates using circuit × year fixed effects (with robust standard errors clustered at the circuit × year level), splitting dissents by panel-majority and panel-minority characteristic, or by age. The sample in column (1) includes only panels with exactly two judges from the same law school, and in columns (2)–(4) includes only mixed-gender panels. p-values test for equality between coefficients in the same column using a Wald test.

Table C.10: Comparing Topic Use Across Panel Types, No Author Controls

Panel type	Variable	(1) Crim.	(2) Civ Rights	(3) 1A	(4) Due Process	(5) Priv.	(6) Lab. Rel.	(7) Econ. & Reg.	(8) Misc.	(9) Unif.	(10) Bayes	Observations
unanimous	Dem.	−0.014	0.031	0.247**	−0.004	−0.000	0.031	−0.064**	0.166***	0.001	0.013*	
		(0.035)	(0.034)	(0.120)	(0.020)	(0.163)	(0.048)	(0.025)	(0.049)	(0.006)	(0.007)	49,098
	Control mean	0.279	0.071	0.006	0.297	0.001	0.057	0.236	0.052	1.753	1.635	
majority-Dem.	3-0 Dem.	−0.001	0.007	0.113	−0.004	0.011	0.030	−0.043**	0.095***	−0.006	0.006	
		(0.023)	(0.024)	(0.074)	(0.013)	(0.108)	(0.033)	(0.018)	(0.034)	(0.004)	(0.005)	61,538
	Control mean	0.257	0.074	0.009	0.279	0.001	0.067	0.247	0.066	1.782	1.660	
majority-Rep.	2-1 Rep.	−0.018	0.007	0.026	0.001	0.038	0.012	0.000	0.036	0.004	0.005	
		(0.017)	(0.016)	(0.059)	(0.010)	(0.079)	(0.020)	(0.012)	(0.024)	(0.003)	(0.003)	101,857
	Control mean	0.279	0.071	0.006	0.297	0.001	0.057	0.236	0.052	1.753	1.635	
bipartisan	Dem. majority	−0.001	0.008	0.080*	0.005	0.009	−0.011	−0.015	0.024	0.003	0.022	
		(0.016)	(0.015)	(0.049)	(0.009)	(0.069)	(0.020)	(0.012)	(0.021)	(0.003)	(0.020)	114,301
	Control mean	0.266	0.072	0.007	0.296	0.001	0.06	0.242	0.055	1.771	1.649	

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Results from conditional logit estimation of predicted topic probabilities (columns (1)–(8)) and Poisson estimation of major topic counts (columns (9) and (10)) using circuit × year fixed effects (with robust standard errors clustered at the circuit × year level), following Equation (3.12). Coefficients represent the effect of making a panel "more Democratic" within the given category (e.g., in the first section, all-Democratic as opposed to all-Republican).

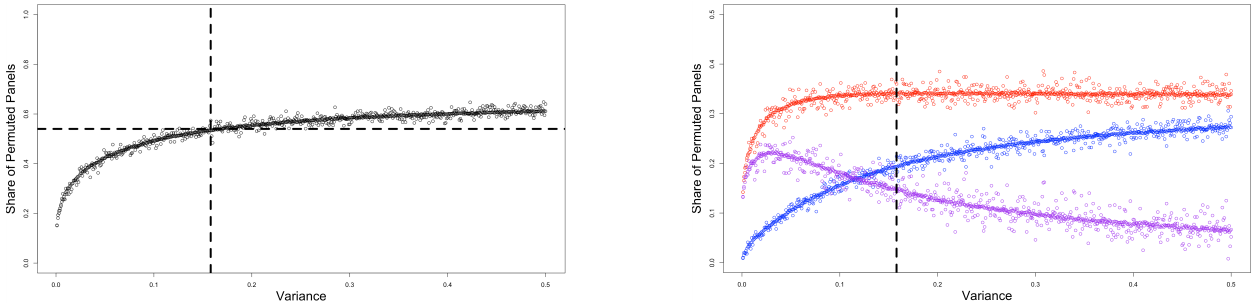
Table C.11: Non-Partisan Characteristics Influence Opinion Similarity

	(1) Topic	(2) Author Party	(3) Panel Majority	(4) # Dems.	(5) Law School	(6) Gender
Normalized match score	0.249***	-0.014***	-0.199***	-0.016***	0.007***	0.015***
	(0.00036)	(0.00022)	(0.00058)	(0.00019)	(0.00005)	(0.00160)
De-meaned match score	0.401***	-0.010***	-0.136***	-0.021***	0.027***	0.004***
	(0.00058)	(0.00015)	(0.00037)	(0.00023)	(0.00018)	(0.00048)
Share de-meaned > 0	0.881	0.438	0.186	0.291	0.682	0.540
Raw match share	0.633	0.495	0.372	0.271	0.074	0.852
Observations	270,451	264,672	246,675	246,675	273,715	274,158

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

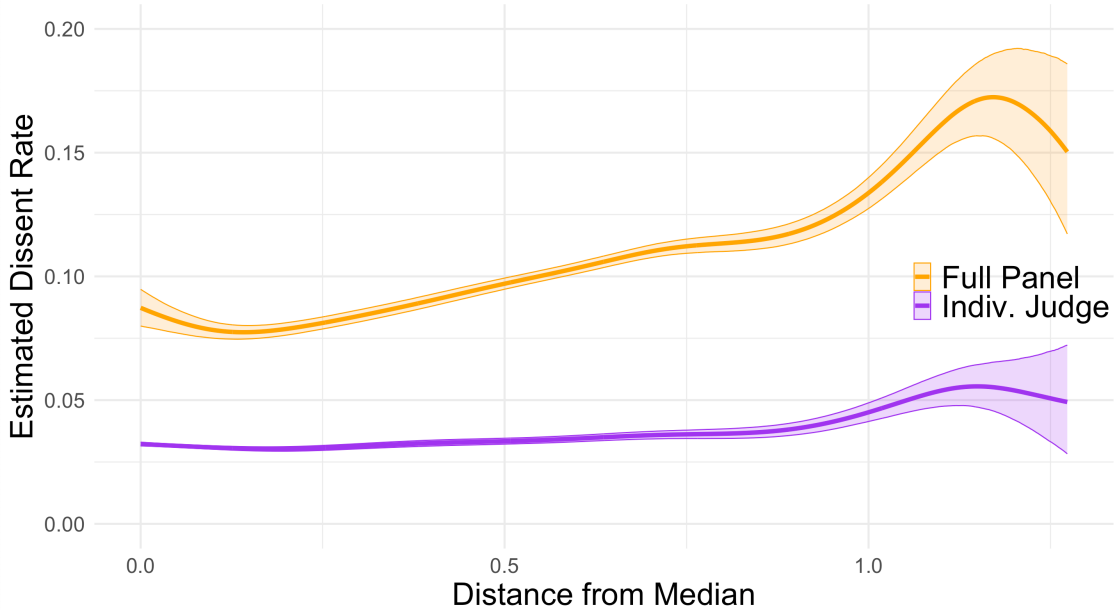
Average shares of 1,000 nearest neighboring opinions matching a given own-opinion characteristic under various normalizations; de-meaning (subtracting the share of opinions matching own characteristics in the full sample) and normalization (de-meaning and division by the entropy of the distribution of characteristics). Standard errors are computed via bootstrap; I draw a new dataset without replacement and re-compute each measure, repeat 10,000 times, and report the standard deviation of the resulting distribution. To calculate statistical difference from 0, I use the share of bootstrap samples with the opposite sign as the observed coefficient, e.g., if 5% of bootstrap samples are weakly positive in column (2), then the coefficient is significant at the $5 \times 2 = 10\%$ level. In fact, no bootstrap draws fit this criterion for any coefficient.

Figure C.1: Shares of Judge Order Permutations As a Function of Noise Variance



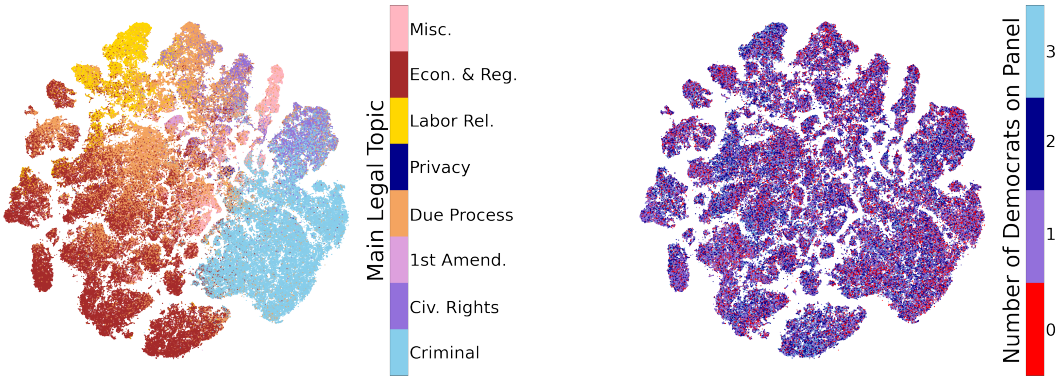
Simulations of judge order with additive i.i.d. $N(0, \sigma^2)$ noise for various values of σ^2. To produce each point, I draw 100,000 panels at random from my data, add independently drawn realizations of $N(0, \sigma^2)$ to each of the judges' JCS scores, then compute the share of post-noise panel draws where the original median, second-extremal, and extremal judges (using original JCS score) are the post-noise median judge. In the left image, I plot the combined share of panels where the original median judge is *not* the post-noise median judge. The dashed vertical line marks the sample variance of JCS scores; the dashed horizontal line marks the share of permuted panels needed to attain the 95% confidence bound for the median judge's dissent rate. In the right panel, I plot the share of panels where the original second-extremal judge is the post-noise median judge (red); the share of panels where the original extremal judge is the post-noise median judge (blue); and the difference in those two shares (purple). The dashed vertical line again marks the sample variance of JCS scores.

Figure C.2: Panel and Individual Dissents as a Function of Distance to Panel Median



Local polynomial regression of dissents as a function of median distance using a local fit of degree 1 and a bandwidth of 0.1. The orange line uses a panel-level specification where the presence of a dissent by any judge on the panel is plotted as a function of the extremal judge's distance from the median. The purple line uses a judge-level specification where each judge's own dissent is plotted as a function of their individual distance from the median. 95% confidence intervals, shown by the shaded areas surrounding each line, are constructed by re-estimating the local polynomial on 10,000 bootstrap samples of the data and plotting the pointwise upper and lower quantiles of those bootstrap results.

Figure C.3: Opinion Embeddings Cluster by Topic, Not by Partisanship

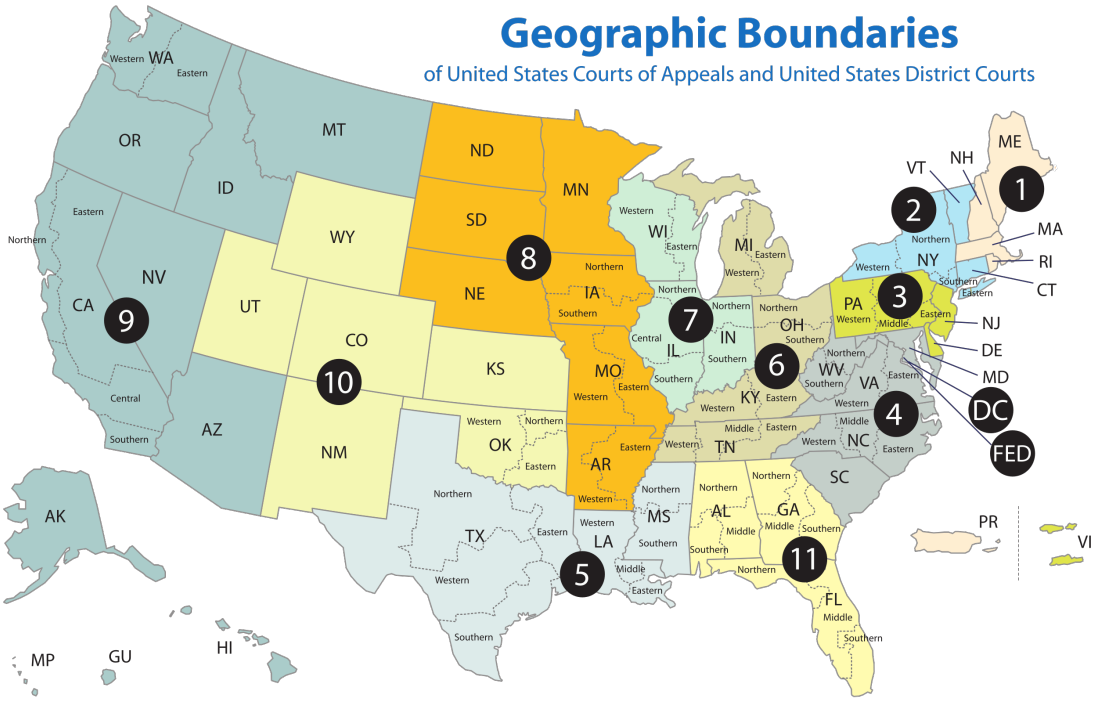


Opinion embeddings visualized using openTSNE in Python. Following package guidelines I use PCA to reduce the 768-dimensional semantic embeddings to 50 dimensions, then apply t-SNE to represent those compressed vectors in two dimensions. I set t-SNE perplexity to 50. The left image colors by main legal topic following the true Bloomberg labels (criminal, civil rights, First Amendment, due process, privacy, labor relations, economic activity and regulation, and miscellaneous). The right image colors by the number of Democrats on the panel (0–3).

C.1 Data Visualizations and Descriptive Statistics

In this appendix, I present additional information about the appellate court data described in Section 3.4. Appendix C.1.1 includes information on opinion data, Appendix C.1.2 includes information on judges and panel compositions, and Appendix C.1.3 includes information on ideology scores. Before those sections, though, I include a figure showing the geographic locations covered by different appellate court circuits.

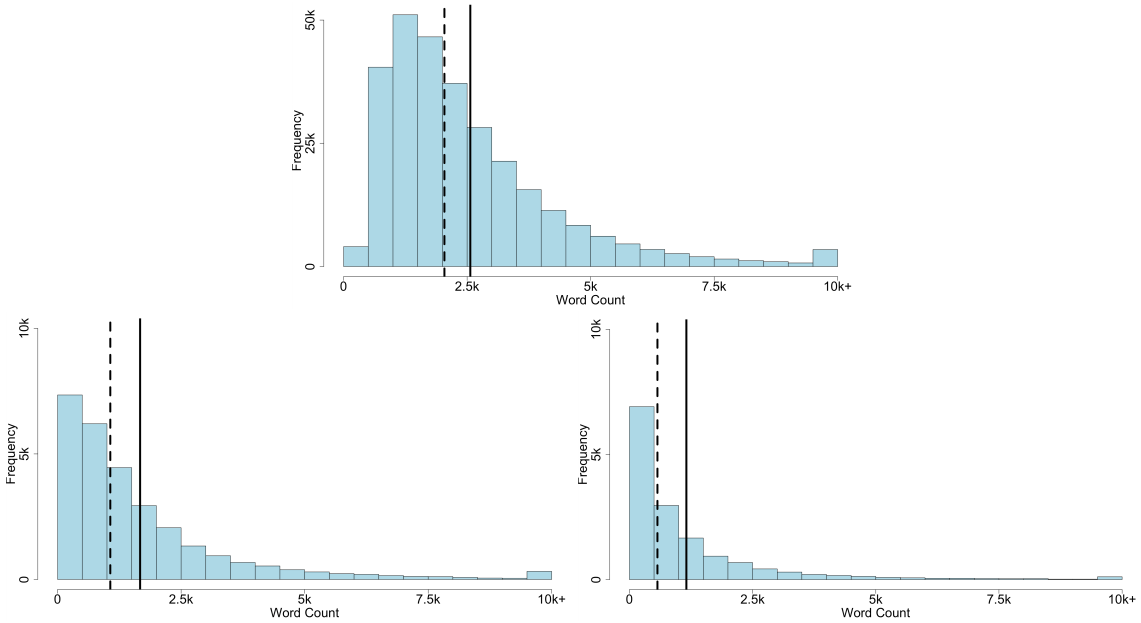
Figure C.4: Official Map of District and Circuit Courts



A map of the thirteen circuit courts of appeals and the district courts they cover. Source: US Federal Government.

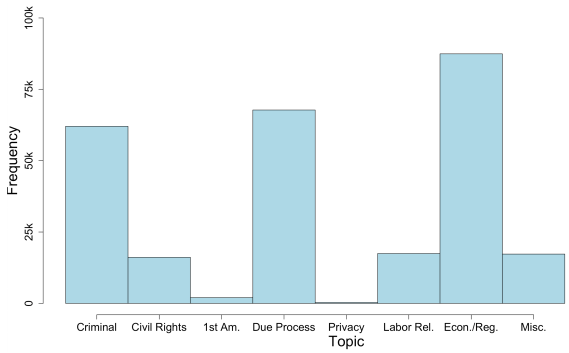
C.1.1 Opinion Characteristics

Figure C.5: Distribution of Word Counts by Opinion Type



In the top center image, word counts for lead/majority opinions where text is available (290,876 opinions); a black solid line marks the mean (2565 words) and a black dashed line marks the median (2401 words). In the bottom left image, the same for dissenting opinions (28,485 opinions; mean 1670 words; median 1067 words). In the bottom right image, the same for concurring opinions (14,879 opinions; mean 1156 words; median 569 words).

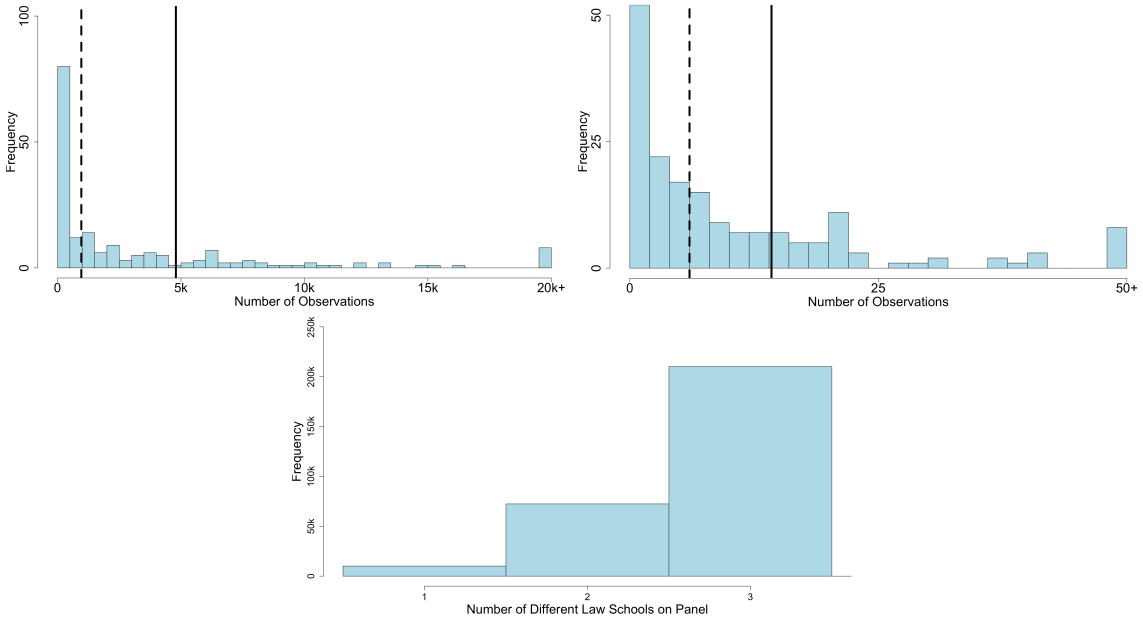
Figure C.6: Distribution of Legal Topics



Primary legal topics for the 270,451 lead opinions where text and topic are available, as coded by Bloomberg Law. Criminal, due process, and economics and regulation cases dominate.

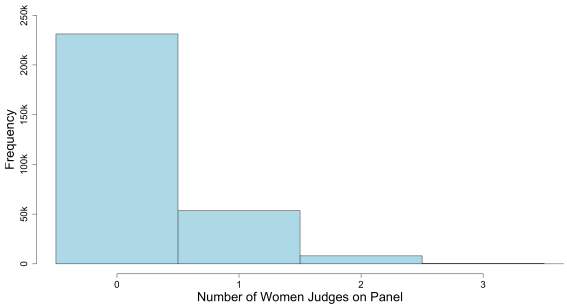
C.1.2 Judge and Panel Characteristics

Figure C.7: Distributions of Judge Law Schools



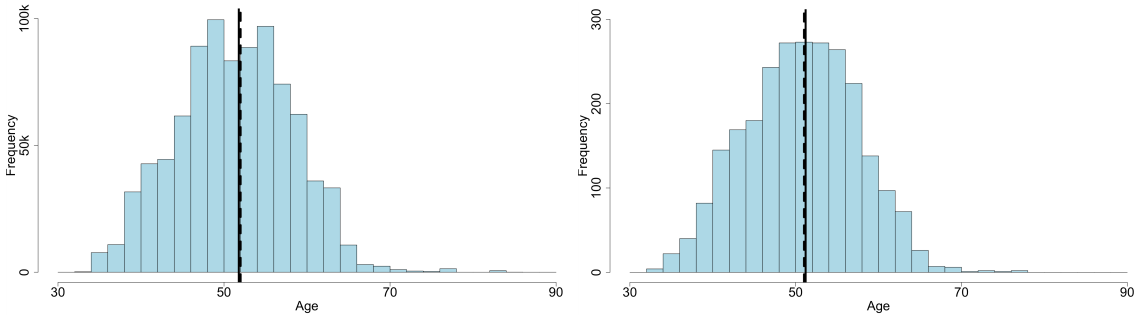
In the top left image, the number of times each of 178 unique law schools appears in the 881,945 observations of law schools on panels with known authors. A black solid line marks the mean (4793), and a black dashed line marks the median (964). In the top right image, the number of times each law school appears across the 2,536 unique judges in that sample (mean 14, median 6). In the bottom center image, the number of panels with known author where one, two, and three different law schools represented (292,636 panels).

Figure C.8: Distribution of Number of Women on Panels



Number of women for the 293,235 panels where author and gender are available. 8.1% of the 882,554 observations of judges and 11.2% of the 2,539 unique judges are women.

Figure C.9: Distribution of Judge Ages



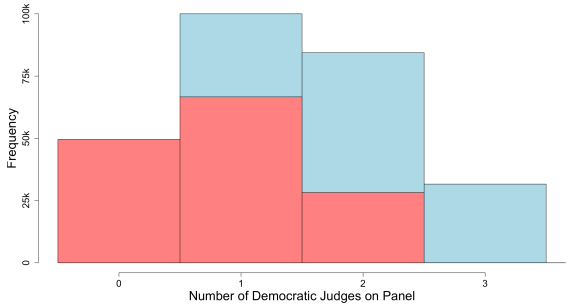
In the left image, judge ages for the 882,554 observations of judges on panels with known authors; a black solid line marks the mean (51.78), and a black dashed line marks the median (52); the standard deviation is 6.95. In the right image, judge ages for the 2,542 unique judges in the sample of panels with known authors (mean 51.20, median 51, standard deviation 6.79).

C.1.3 Judge Ideology Scores

I first describe in more detail the three measures of judge ideology used in this work, then present visualizations of those measures.

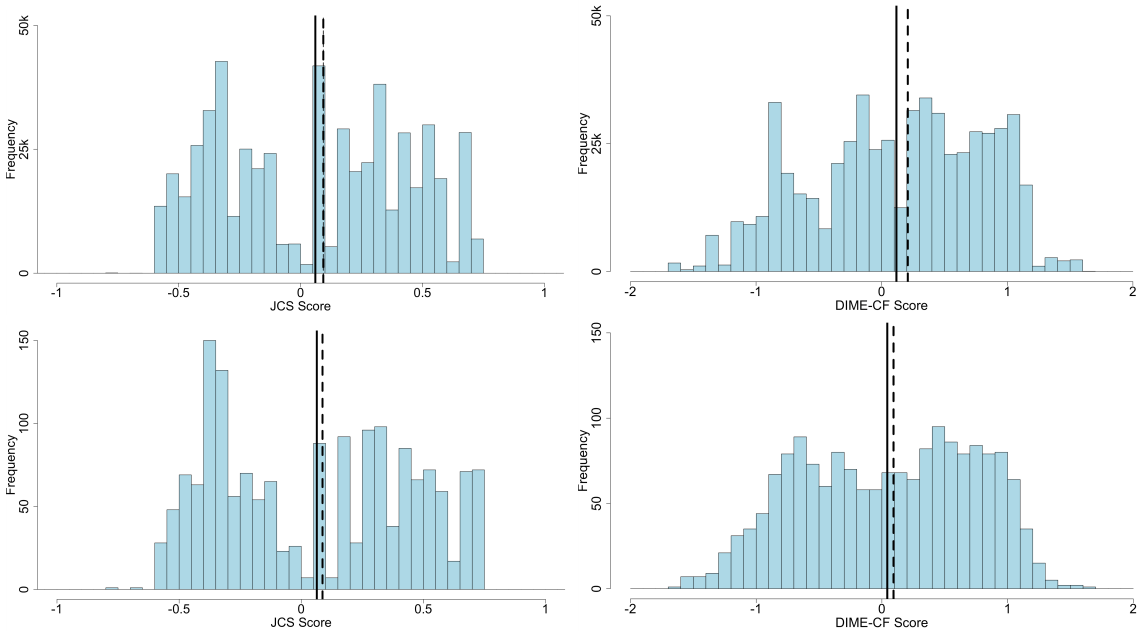
- Self-reported party; available for almost all judges. This value is coarse but model-free. In all but 46 of the over one million observations, this value matches the party of the appointing president (another standard measure in the literature).
- Judicial Common Space (JCS) scores, from Epstein et al. [35]; available from 1970 onward. These scores are based on home-state senator influence in appellate appointments (see Section 3.3.1). JCS scores average the DW-Nominate score of the appointing president with those of home-state senators in the president's party. DW-Nominate scores are from Poole and Rosenthal [66], who estimate a structural model of Congressional votes; legislators who vote similarly have similar ideal points.
- Database on Ideology, Money in Politics, and Elections—Campaign Finance (DIME-CF) scores, from Bonica and Sen [16]; available from 1970 onward. These scores are based on linking judges (and other officials) to their political contributions and applying ideal-point estimation in the style of DW-Nominate. Correlation with JCS and DW-Nominate scores is high, but the authors claim DIME-CF scores better capture judge preferences rather than political deals in the appointment process.

Figure C.10: Distribution of Number of Democrats on Panels



Number of Democrats on a panel for the 265,592 panels where author and panel parties are available. Stacked blue and red bars represent the share of opinions authored by Democrats and Republicans, respectively. 45.2% of the 834,153 observations of judges, and 45.3% of the 2,407 unique judges, in this sample are Democrats.

Figure C.11: Distributions of Judge Ideology Scores



In the top images, JCS and DIME-CF scores for the 547,738 and 554,930 observations, respectively, of those scores on panels with known authors. A black solid line marks the mean, and a black dashed line marks the median. Observed JCS scores are roughly bimodal (mean 0.060, variance 0.148), while observed DIME-CF scores are roughly uniform (mean 0.117, variance 0.468). The bottom images replicate the analysis for unique judges (samples of 1,682 and 1,698 respectively); JCS scores are again roughly bimodal (mean 0.064, variance 0.158), while DIME-CF scores now appear bimodal (mean 0.043, variance 0.478).

C.2 Quasi-Random Assignment of Judges

In Appendix C.2.1, balance tables C.12 – C.14 aggregate across the full sample to show that the share of cases in each of the eight Bloomberg Law topics is relatively consistent across panel types, and that law school and gender composition is roughly orthogonal to partisanship. Next, in Appendix C.2.2, I generate—using 1,000 Bayesian bootstrap draws of the set of observed panels—distributions of the shares of panels with 0–3 Democrats; 1–3 unique law schools; and 0–3 women. Those distributions are shown in figures C.12–C.14. Finally, in Appendix C.2.3, I describe and implement a rigorous procedure to test for random assignment at the individual circuit × year level; results are in figures C.15–C.17.

C.2.1 Balance Tables for Topics and Panel Characteristics

Looking first to demographics (tables C.12 and C.13), the balance tables show a slightly higher probability for women judges to be Democrats, and for panels with judges from three different law schools to have larger numbers of Democrats. Since there are more panels in the sample with a Democratic judge in the minority than with a Republican judge in the minority, these findings translate into a mild negative correlation between bipartisan panels and panels with judges of three different law schools, as well as a mild positive correlation between bipartisan panels and panels with no women. In the framework of Section 3.5.3, the former slightly weakens the degree to which law school must outweigh party, since double-minority judges across party and law school are more rare than if those characteristics were entirely orthogonal. By the same argument, it slightly strengthens the degree to which gender must outweigh party.

In Table C.14, the share of cases in each major topic (criminal, due process, economics and regulation) vary only slightly with panel composition. Panels with more Democrats have slightly fewer cases in the former two topics and slightly more in the latter; the same is true of panels with fewer unique law schools. Panels with more women have more criminal and due process cases and fewer economics and regulation cases. These topic codes take into account both the facts of the case (i.e., pre-trial information) and the ultimate opinion, which may account for some of these trends. Importantly, balance is much stronger when comparing bipartisan to party-unanimous panels, as well as comparing majority-Democrat to majority-Republican panels. These results suggest that differences in case topics are not driving differences in aggregate panel behavior.

Table C.12: Distribution of Panels by Number of Law Schools

	Panels w/o Missing Law School	3 Diff. Law Schools	2 Diff. Law Schools	1 Law School
Full sample	292636	0.717	0.248	0.035
Sample w/o missing party	264,993	0.712	0.252	0.036
Panels w/ 0 Dems.	49,323	0.693	0.264	0.043
Panels w/ 1 Dem.	99,684	0.703	0.249	0.048
Panels w/ 2 Dems.	84,365	0.720	0.253	0.027
Panels w/ 3 Dems.	31,621	0.756	0.235	0.009
Bipartisan panels	184,049	0.711	0.251	0.038
Party-unanimous panels	80,944	0.717	0.253	0.030
Majority-Dem. panels	115,986	0.730	0.248	0.022
Majority-Rep. panels	149,007	0.700	0.254	0.046
Sample w/o missing gender	292,636	0.717	0.248	0.035
Panels w/ 0 women	230,580	0.694	0.264	0.042
Panels w/ 1 women	53,550	0.798	0.194	0.008
Panels w/ 2 women	8,037	0.847	0.150	0.003
Panels w/ 3 women	469	0.889	0.111	0

Table C.13: Distribution of Panels by Number of Women

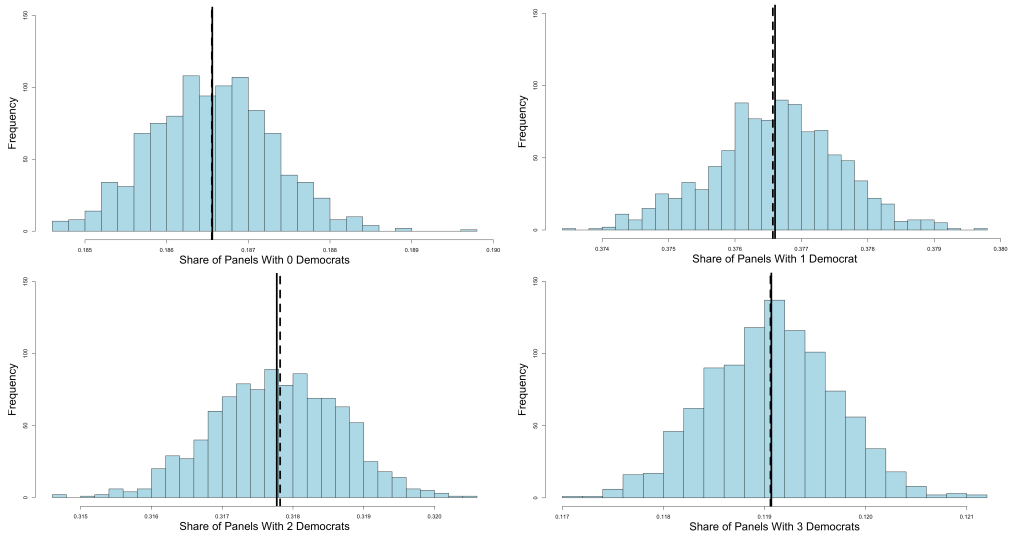
	Panels w/o Missing Gender	0 Women	1 Woman	2 Women	3 Women
Full sample	292,636	0.79	0.183	0.027	0.002
Sample w/o missing party	264,993	0.816	0.163	0.022	0.001
Panels w/ 0 Dems.	49,323	0.827	0.161	0.017	0.001
Panels w/ 1 Dem.	99,684	0.802	0.177	0.023	0.001
Panels w/ 2 Dems.	84,365	0.811	0.164	0.025	0.001
Panels w/ 3 Dems.	31,621	0.861	0.118	0.02	0.001
Bipartisan panels	184,049	0.806	0.171	0.024	0.001
Party-unanimous panels	80,944	0.840	0.144	0.018	0.001
Majority-Dem. panels	115,986	0.824	0.151	0.024	0.001
Majority-Rep. panels	149,007	0.810	0.171	0.021	0.001

Table C.14: Distribution of Case Topics Across Panel Types

	Cases w/o missing topic	Crim.	Civ. Rights	1st Am.	Due Process	Priv.	Labor Rel.	Econ. & Reg.	Misc.
Full sample	289,894	0.231	0.060	0.007	0.249	0.001	0.064	0.323	0.064
Sample w/o missing party	260,829	0.227	0.059	0.007	0.245	0.001	0.064	0.335	0.062
Panels w/ 0 Dems.	48,932	0.251	0.063	0.006	0.262	0.001	0.057	0.297	0.063
Panels w/ 1 Dem.	98,043	0.232	0.059	0.006	0.252	0.001	0.060	0.327	0.062
Panels w/ 2 Dems.	82,953	0.218	0.058	0.007	0.236	0.001	0.068	0.349	0.062
Panels w/ 3 Dems.	30,901	0.198	0.053	0.008	0.221	0.001	0.077	0.379	0.063
Bipartisan panels	180,996	0.226	0.059	0.007	0.245	0.001	0.064	0.337	0.062
Party-unanimous panels	79,833	0.230	0.059	0.007	0.246	0.001	0.065	0.329	0.063
Majority-Dem. panels	113,854	0.213	0.057	0.007	0.232	0.001	0.070	0.357	0.062
Majority-Rep. panels	146,975	0.238	0.060	0.006	0.255	0.001	0.059	0.317	0.062
Sample w/o missing gender	288,404	0.231	0.06	0.007	0.249	0.001	0.064	0.323	0.064
Panels w/ 0 women	226,553	0.224	0.056	0.006	0.235	0.001	0.065	0.353	0.059
Panels w/ 1 woman	53,375	0.257	0.077	0.010	0.300	0.002	0.059	0.219	0.077
Panels w/ 2 women	8,008	0.257	0.077	0.011	0.316	0.003	0.054	0.185	0.097
Panels w/ 3 women	468	0.250	0.085	0.009	0.297	0.002	0.051	0.203	0.103
Sample w/o missing law school	287,823	0.231	0.060	0.007	0.250	0.001	0.064	0.323	0.064
Panels w/ 3 diff. law schools	206,650	0.241	0.065	0.008	0.260	0.001	0.067	0.293	0.065
Panels w/ 2 diff. law schools	71,286	0.214	0.053	0.006	0.229	0.001	0.058	0.378	0.061
Panels w/ 1 law school	9,887	0.144	0.030	0.003	0.188	0.001	0.032	0.548	0.055

C.2.2 Panel-Level Simulations with Aggregate Data

Figure C.12: Distributions of Aggregate Shares of Panel Partisanship



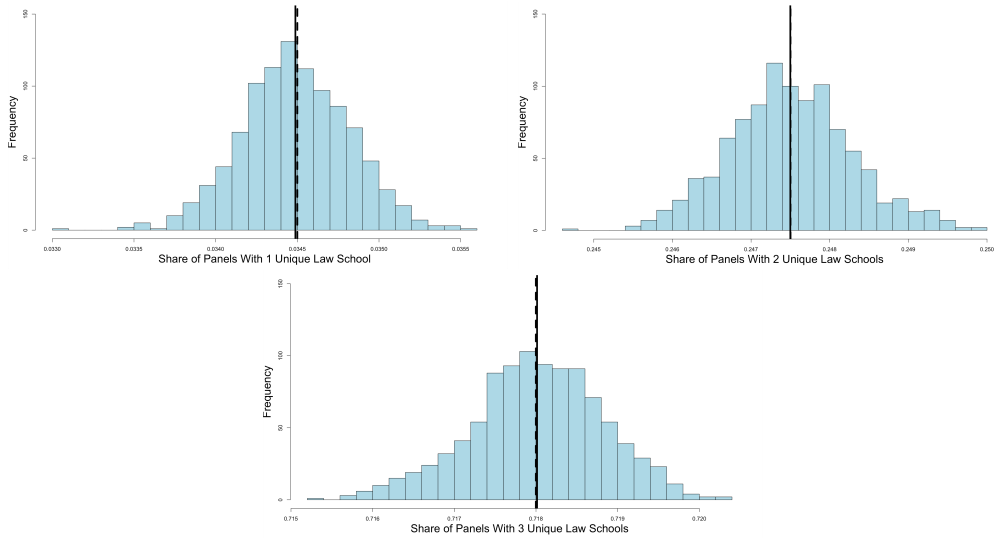
Shares of panels with 0–3 Democrats, sampled at the panel level. The bootstrap means (solid lines) are 0.1866, 0.3766, 0.3178, and 0.1191 respectively; the observed shares (dashed lines) are 0.1866, 0.3766, 0.3178, and 0.1191 respectively.

C.2.3 Judge-Level Simulations by Circuit times Year

Aggregate balance checks may mask non-random assignment within particular circuit × year pairs ct. To address this concern, I describe a procedure for sampling from the distribution of judges in each ct, using number of Democrats on a panel as an example. (It is the same for number of unique law schools and women.) For each ct, I do the following:

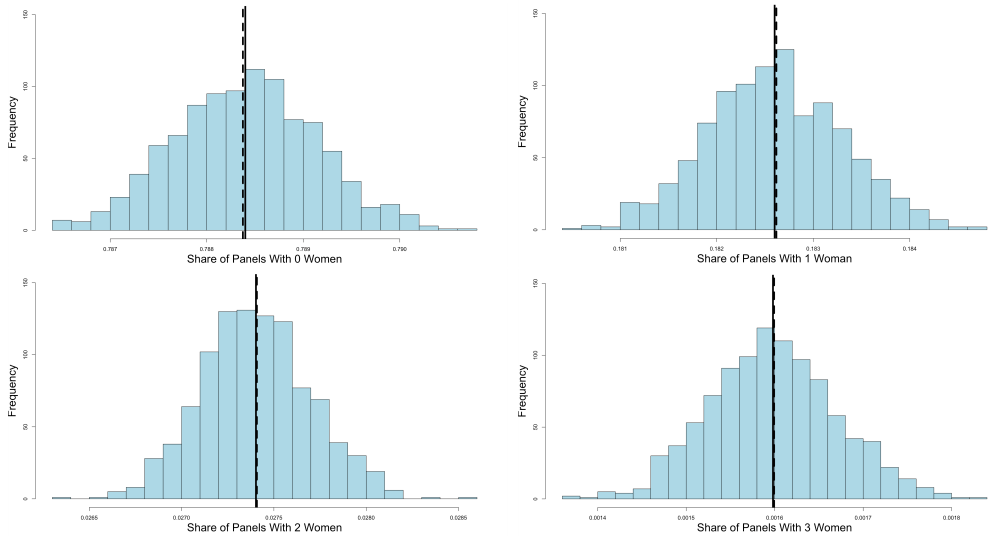
1. Compute the distribution F_{ct} of appearances by unique judges in the N panels on ct.
2. Draw a panel p_1 of 3 judges without replacement from F_{ct}. Draw N independent panels $p_1, ..., p_N$ (replacing judges between panel draws) to form a simulated ct, s^1_{ct}.
3. Draw 1,000 independent simulations $s^1_{ct}, ...s^{1000}_{ct}$. For each s^i_{ct}, compute
   - the share of those n panels with 0, 1, 2, and 3 Democratic judges.
   - the share of those n panels with 1, 2, and 3 unique law schools.
   - the share of those n panels with 0, 1, 2, and 3 women.
4. For each of those 11 shares, form an empirical distribution G using the s^i_{ct}.

Figure C.13: Distributions of Aggregate Shares of Panel Law Schools



Shares of panels with 1–3 unique law schools, sampled at the panel level. The bootstrap means (solid lines) are 0.0345, 0.2475, and 0.7180 respectively; the observed shares (dashed lines) are 0.0345, 0.2475, and 0.7180 respectively.

Figure C.14: Distributions of Aggregate Shares of Panel Genders

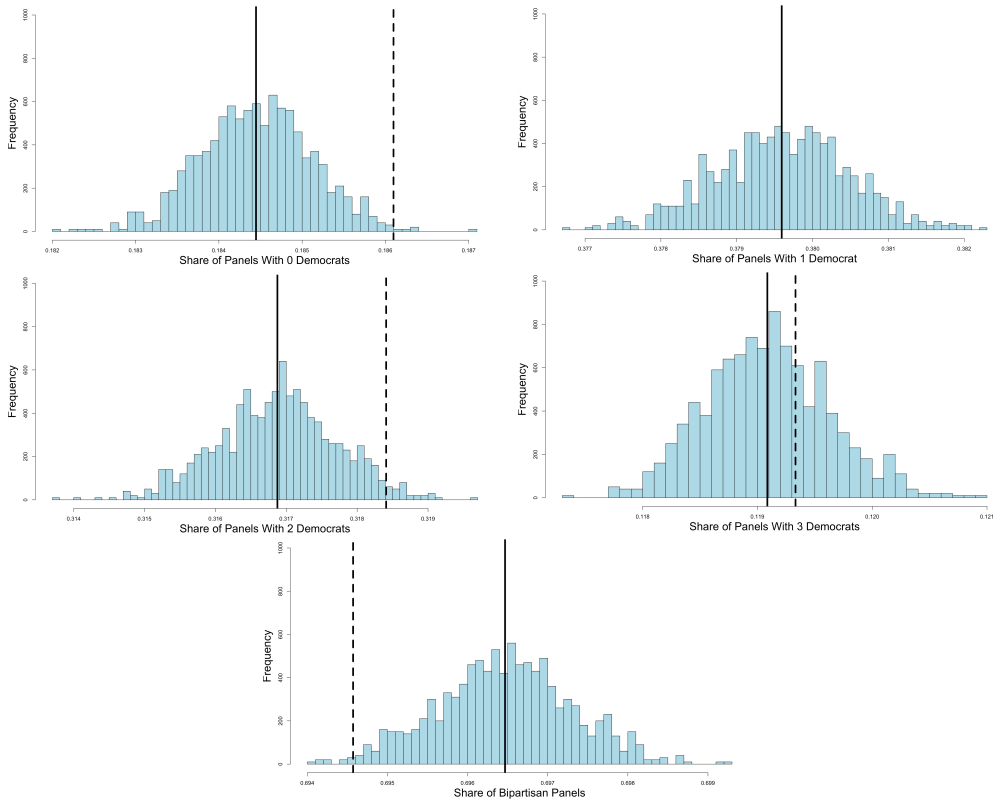


Shares of panels with 0–3 women, sampled at the panel level. The bootstrap means (solid lines) are 0.7884, 0.1826 0.0274, and 0.0016 respectively; the observed shares (dashed lines) are 0.7884, 0.1826, 0.0274, and 0.0016 respectively.

The procedure can be used to return a list of all ct where the observed quantile indicates statistically significant deviations from random assignment; below I report the aggregate shares across all ct. It requires hundreds of millions of random draws, but does not consider case topics or joint distributions; e.g., panels with 2 Democratic judges, one of whom is a woman and shares a law school with the Republican man. Since these tests are already much more thorough than any previous economics literature and show only small deviations from random assignment, I rely on the aggregate balance checks to test these cases.

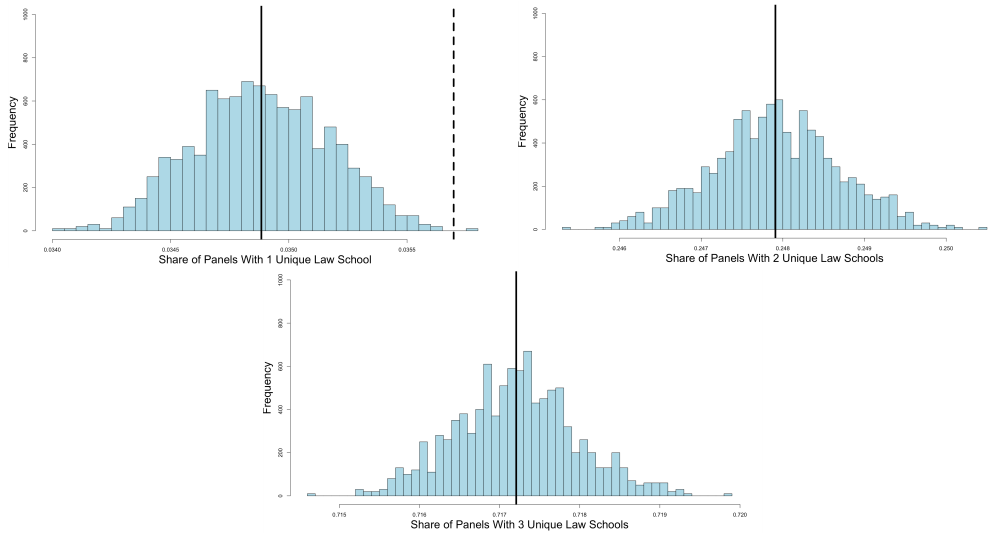
Figures C.15–C.17 show some statistically significant deviations from random assignment, but the magnitudes are small and sensitive to data with missing observations. No share of Democrats or unique law schools differs by more than 0.5 percentage points from the simulated mean; differences for shares of women judges are slightly larger but driven almost entirely by a low observed share of panels with two women judges.

Figure C.15: Distributions of Aggregate Shares of Panel Partisanship, Judge-Level Simulations



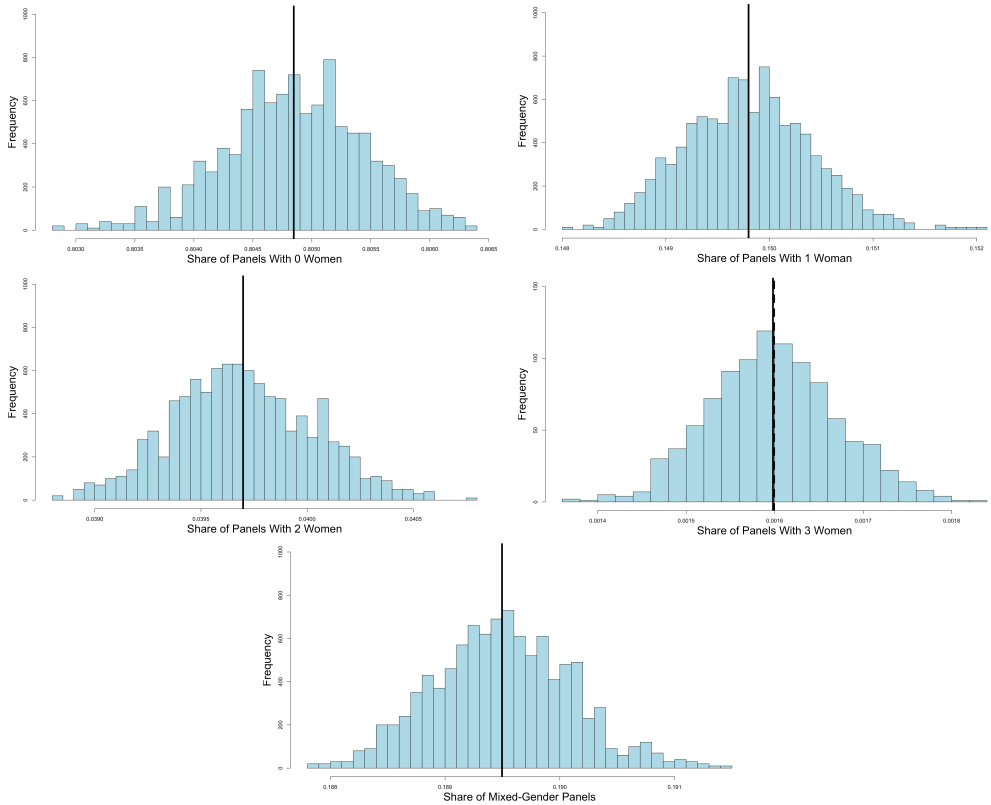
Shares of panels with 0–3 Democrats, sampled at the judge level. The bootstrap means (solid lines) are 0.1845, 0.3780, 0.3169, 0.1191, and 0.6965 respectively; the observed shares (dashed lines, when visible) are 0.1861, 0.3762, 0.3184, 0.1193, and 0.6946 respectively.

Figure C.16: Distributions of Aggregate Shares of Panel Law Schools, Judge-Level Simulations



Shares of panels with 1–3 unique law schools, sampled at the judge level. The bootstrap means (solid lines) are 0.7172, 0.2479, and 0.0345 respectively; the observed shares (dashed lines, when visible) are 0.7128, 0.2515, and 0.0357 respectively.

Figure C.17: Distributions of Aggregate Shares of Panel Genders, Judge-Level Simulations



Shares of panels with 0–3 women, sampled at the judge level. The bootstrap means (solid lines) are 0.8048, 0.1498, 0.0397, 0.0057, and 0.1895 respectively; the observed shares (dashed lines, when visible) are 0.8144, 0.1624, 0.0221, 0.0012, and 0.1845 respectively.

C.3 Theoretical Appendix

In this section, I describe and prove some additional results for the model in Section 3.5.1. Appendix C.3.1 proves results for sincere dissents with noise; Appendix C.3.2 extends that model by adding a role for opinions as well as dispositional votes; and Appendix C.3.3 provides details of the coalitional similarity framework with multidimensional judges.

C.3.1 Additional Results for Sincere Dissents with Noise

I add case-specific noise to the model of Section 3.5.1, representing the fact that a judge's cutoff for upholding a ruling may vary depending on the precise facts at hand. For each case c, judges experience i.i.d. preference shocks $\eta_j \sim N(0, \sigma^2)$, and then vote sincerely based on their new cutoffs $x_j + \eta_j$. Each judge's cutoff is thus a random variable $x'_j \sim N(x_j, \sigma^2)$.

Despite the relative tractability of the normal distribution, it is immediately clear why an exact analytic expression for dissent probabilities is not possible. By additivity, the distance between judges j and k is a normally distributed random variable D_{jk}, as is that distribution conditional on some fixed distance δ between judges i and k, $D_{jk}|\delta$. Judge j may dissent either when $D_{jk} < 0$ and $D_{ik} < 0$, or $D_{jk} > 0$ and $D_{ik} > 0$. Both of these expressions involve the product of a normal distribution with the integral of the normal cdf and have no closed form solution. Furthermore, the resulting expression would not have uniform density over the interval of possible distances. An easier result to establish is the claim that there is a larger increase in total dissents from increasing the extremal distance than the second-extremal distance:

Proposition 10. *In the sincere dissents model with i.i.d. $N(0, \sigma^2)$ noise, there exists a constant K such that for all $\sigma^2 < K$, increasing the second-extremal distance leads to a smaller increase in total dissents than increasing the extremal distance.*

Proof. I prove the proposition in three steps. First, with sufficiently small noise, only three of the six possible permutations of judge orders need be considered. Next, I show that the second-extremal judge is more likely to be permuted to the median than the extremal judge. Last, I show that this effect directly produces a lower effect of varying second-extremal distance.

Consider the permutation $x'_2 < x'_1 < x'_3$, where judge 1 becomes the median judge. To attain it, it is at least necessary that the realization of D_{12} is sufficiently large to produce $x'_2 < x'_1$. By choosing σ^2 small enough, such a realization lies in the tail of the (normally distributed) D_{12}. The probability of a permutation that is twice as large—which would be needed for $x'_3 < x'_1$—is thus approximately zero in comparison. I can therefore rule out the permutations $x'_2 < x'_3 < x'_1$, $x'_3 < x'_2 < x'_1$, and $x'_3 < x'_1 < x'_2$, which proves the first claim.

Next, I show that of the three remaining permutations, the one which swaps judges 1 and 2 and leaves at least ε distance between them is more likely than the one which swaps judges 2 and 3 and leaves at least ε distance between them. By assumption, the distance between judges 1 and 2 is smaller than the distance between judges 2 and 3. Thus, a realization of D_{12} which makes $x'_2 < x'_1$ is more likely than a realization of D_{23} which makes $x'_3 < x'_2$. Indeed, for σ^2 sufficiently small both realizations are in the tails of their respective distributions (which

have the same variances) and thus the probability of the second realization is exponentially smaller than the probability of the first.

Finally, note that when comparing a realized value $x_1 = x$ to a smaller realization $x_1 = x - \varepsilon$, total dissents will increase under the second realization unless $x'_1 - x'_2 > \varepsilon$. Likewise, when comparing $x_3 = y$ to $x_3 = y + \varepsilon$, total dissents will increase unless $x'_2 - x'_3 > \varepsilon$. As established above, the first event is more likely than the second, so total dissents are more likely to increase when x_3 is increased by ε than when x_1 is decreased by ε. □

For $\sigma^2 = 0$ there is no difference in the increase in dissent rates caused by increasing second-extremal and extremal distances. As σ^2 becomes arbitrarily large, the original mean x_j is irrelevant and all permuations of judges become equally likely. Thus there is an important non-monotonicity in the relative importance of extremal distance with respect to noise. In the right image of Figure C.1, I show that both the probability of permuting the second-extremal judge to the median and that of permuting the extremal judge to the median rise with σ^2, but the difference rises for $\sigma^2 \in [0, 0.03]$ before falling for $\sigma^2 \in [0.03, 0.5]$. While a full theoretical characterization of this behavior is beyond the scope of this work, the simulated results in Section 3.5.2 provide some suggestive evidence that such non-monotonicities are relevant in real-world contexts, even if they do not ultimately affect the fit of this model with my leading empirical results.

C.3.2 Strategic Dissents

To connect the sincere voting setup to classical median voter models, I consider a simplification of Parameswaran et al. [65] and introduce a policy utility term that captures not only the disposition of a case, but also the location of the opinion issued by the dispositional majority. Let the location of an opinion o be equal to the median of the x_j among judges j in the dispositional majority. Judges face a policy cost $p(x_j, 0) = p(|x_j - o|) = p(d)$, reflecting the fact that such an opinion shapes future policy in ways that a judge dislikes more the further that opinion is from their ideal point. To introduce a reason for judges to vote strategically on the disposition, I also impose an expressive cost $q(x_j, c) = q(|x_j - c|) = q(d)$ for a judge j who votes against their preferred disposition on case c. By making these costs a function of the distance between x_j and c, I automatically impose location invariance and symmetry, i.e., x_j near zero and x_j near 1 are treated the same, as are insincere upholdings and insincere reversals. I also assume that $p(0) = q(0) = 0$,[1] that both costs are continuous, and that both costs are weakly increasing in $|x_j - o|$ and $|x_j - c|$ respectively.

If judge 1 or judge 3 are part of a unanimous opinion and choose to dissent, the opinion will be further from their preferred x_j, since it will be located at the average of the remaining two judges' ideal points instead of being located at the median judge's ideal point. On a fixed panel, this cost does not depend on c, but will differ for the two judges. My next result establishes how the cost of dissent varies with the location of the extremal judges.

Proposition 11. *The cost of dissent to judge $i \in \{1, 3\}$ decreases with the distance of judge i from the median if and only if the policy cost $p(x_i, o)$ has increasing differences in (x_i, o).*

[1] Generically, $x_1 \neq x_2 \neq x_3$ so that $|x_j - o| = 0$ is never relevant for judges 1 and 3.

The policy cost to that same judge increases with the distance of judge i from the median if and only if it has decreasing differences in (x_i, o).

Proof. The definition of increasing differences requires

$$x_i \geq x_i' \text{ and } o \geq o' \Rightarrow p(x_i, o) - p(x_i', o) \geq p(x_i, o') - p(x_i', o').$$

The cost of dissent to judge i when their ideal point is x_i is

$$p(x_i, o) - p(x_i, o'),$$

where o is the location of the opinion in the coalition with judge i and o' is the location of the opinion in the coalition without judge i.

Letting $i = 1$, I have $o > o'$, since without judge 1 the opinion is set at the median of judges 2 and 3 rather than at the ideal point of judge 2. Moving judge 1 further from the median requires $x_1' < x_1$, since judge 1 is to the left of the median. Therefore the condition for a decreasing cost of dissent is

$$p(x_1', o) - p(x_1', o') \leq p(x_1, o) - p(x_1, o') \Leftrightarrow p(x_1, o') - p(x_1', o') \leq p(x_1, o) - p(x_1', o),$$

which is precisely the expression for increasing differences. Likewise, an increasing cost of dissent requires the opposite inequality, which is precisely the condition for decreasing differences.

Letting $i = 3$, I instead have $o < o'$ and $x_1' > x_1$. Switching both inequalities means that once again the condition for a decreasing cost of dissent is precisely the expression for increasing differences and the condition for a decreasing cost of dissent is precisely the expression for decreasing differences. □

One simple example of a policy cost with increasing differences satisfying the earlier conditions is a logarithmic loss $p(x_i, o) = \ln[(x-y)^2]$; since this expression is twice continuously differentiable away from 0, I can check that its cross-partial derivative is $2/(x-y)^2 > 0$. A simple policy cost with decreasing differences satisfying the earlier conditions is the quadratic loss $p(x_i, o) = (x_i - o)^2$; since this expression is twice continuously differentiable, I can check that its cross-partial is $-2 < 0$.

This result, combined with symmetry of the policy cost, immediately establishes that under increasing (decreasing) differences the cost of dissent is higher (lower) for judge 1 than for judge 3 because judge 1 is closer to the median than judge 3. Letting δ_1 be the cost of dissent to judge 1 and δ_3 be the cost of dissent to judge 3, I can now establish results about the relative dissent rates of the two judges.

Proposition 12. *If $\delta_1 \geq \delta_3$, then extremal judges dissent weakly more often. The difference is strict if extremal judges dissent at all.*

Proof. I begin with the case $\delta_1 = \delta_3 = \delta$. Because $q(d)$ is weakly increasing and symmetric, it attains its maximum on $[0, x_2 - x_1]$ at $d_{12} = x_2 - x_1$ and attains its maximum on $[0, x_3 - x_2]$ at $d_{23} = x_3 - x_2$. Furthermore, $q(d_{12}) \leq q(d_{23})$. If $\delta \geq q(d_{23})$, then the maximal expressive cost of insincerity is lower than the common cost of dissent, so neither judge ever dissents.

If $q(d_{12}) \leq \delta < q(d_{23})$, only judge 3 may dissent. The remaining case is therefore where $q(d_{12}) < \delta$.

Since $q(0) = 0 < \delta$, there is a well-defined closed interval $I = [x_\ell, x_u] \subset (0, x_2 - x_1)$ where $d = q(x_2 - x') \,\forall\, x' \in I$. (Of course, if $q(d)$ is strictly increasing, $x_\ell = x_u$ and that interval is a point). Judge 1 will thus insincerely join the majority whenever $c \in [x_1, x_1 + x_\ell]$. This interval is not empty, but may be the entirety of $[x_1, x_2]$.

By symmetry and equal costs of dissent, judge 3 will insincerely join the majority on exactly the same measure of cases as judge 1. However, judge 3 faces a strictly larger interval of cases where they sincerely disagree with the majority, and will thus dissent more often than judge 1.

Now consider the case $\delta_1 > \delta_3$. When $q(d_{23}) < \delta_3$, the maximal cost of insincerity is lower than the minimal cost of dissent so neither judge ever dissents. When $\delta_3 < q(d_{23})$ but $q(d_{12}) < \delta_1$ (regardless of the relationship between $q(d_{23})$ and δ_1, and the relationship between $q(d_{12})$ and δ_3), the maximal cost of insincerity remains lower than the cost of dissent for judge 1, who never dissents, but judge 3 may sometimes dissent. The remaining case is $\delta_1 < q(d_{12}) \leq q(d_{23})$, in which case both judges may dissent.

By the same argument above, there are cutoffs d^*_{23} and d^*_{12} such that judge 3 will insincerely join the majority only for cases closer than d^*_{23} to x_3 and judge 1 will insincerely join the majority only for cases closer than d^*_{12} to x_1. However, because $\delta_1 > \delta_3$ and both judges face the same expressive cost, it must be that $d^*_{23} < d^*_{12}$, i.e., judge 1 insincerely joins the majority for a larger measure of cases than judge 3.

Since judge 3 faces both a larger measure of cases where dissent minimizes their expressive cost, and a smaller measure of cases where they insincerely join the majority, the extremal judge dissents more often. □

The expected converse result—that when $\delta_1 < \delta_3$, second-extremal judges dissent more often—does not hold because of two offsetting effects. As seen in the proof, a higher cost of dissent means a larger measure of cases where the judge insincerely joins the majority; however, the total measure of cases where judge 3's expressive preferences predict dissent is larger than the measure of cases where judge 1's expressive preferences predict dissent. Thus, the key question is whether the cost of dissent grows more than one-to-one with distance from a judge's ideal point. If this property holds, the higher cost of dissent for judge 3 outweighs the larger share of cases where their expressive preferences predict dissent (since that share grows one-to-one with distance because of the uniform distribution of cases). Thus, judge 3 actually dissents less often than judge 1.

That same property is important in understanding the effect of increasing second-extremal and extremal distances from the median. If the cost of dissent has increasing differences, then the cost of dissent is decreasing with distance. Moving either judge 1 or judge 3 further from the median increases the share of cases where expressive preferences predict a dissent and the share of cases where the cost of dissent is low enough to permit a sincere dissent. Thus the total share of dissents by each judge increases. That share is higher for the extremal judge than the second-extremal judge; whether the increase is more rapid for one judge than the other depends on the curvature of the cost of dissent.

If instead the cost of dissent has decreasing differences, then moving either judge 1 or judge 3 further from the median increases the share of cases where expressive preferences

predict a dissent but decreases the share of cases where the cost of dissent is low enough to permit a sincere dissent. In this case, the curvature of the cost of dissent determines both whether the amount of dissents increases and for which judge it changes more quickly. With a convex cost, the increase in insincere agreements outweighs the increase in cases with possible dissents; this effect is stronger for the extremal judge, who dissents less often than the second-extremal judge, and leads to an extremal dissent rate that falls more quickly than the second-extremal dissent rate. With a concave cost, dissents increase for both judges, though more slowly than if the cost of dissent had increasing differences because the increase in cases with possible dissents must compete with the increase in insincere agreements. Since cost is concave, the extremal judge dissents more often than the second-extremal judge, and the growth in their dissent rate is faster than that of the second-extremal judge.

This model shows the complex possibilities of a fully strategic framework modeling both case dispositions and opinion locations, but falls short of capturing the empirical facts observed in the appellate court data. In particular, dissents increase when the extremal judge moves further from the median, suggesting a world in which either the cost of dissent has increasing differences and is convex or the cost of dissent has decreasing differences and is concave. But in both of these settings, the extremal judge dissents more often than the second-extremal judge, and the median judge does not dissent at all. As shown in the sincere dissents framework, the presence of noise can dramatically complicate results, and is beyond the scope of this analysis. Even still, the simulation results in Section 3.5.2 suggest that noise alone is insufficient to induce equal dissent rates by median and extremal judges; more exploration of this rich framework is needed to arrive at a strategic justification for the empirical fact.

C.3.3 Multidimensional Judges and Coalition Formation

Let each judge have an N-dimensional type $t_i = (x_i^1, ...x_i^N)$, where the first dimension captures partisanship and the remaining dimensions capture other characteristics, e.g., demographic features such as gender and age or non-partisan ideological features such as law school training or previous clerkships.

An *opinion* is a partition of the judges into two groups called *coalitions*,[2] one of which may be empty. For example, $(\{1,2,3\},\{\emptyset\})$ represents a unanimous opinion, while $(\{1,2\},\{3\})$ represents a split opinion with judges 1 and 2 in the majority and judge 3 dissenting. The value of any coalition C with at least two judges is given by

$$\frac{1}{|C|}\sum_{\{i,j\}\in C}\langle t_i, t_j\rangle,$$

the average weighted dot product of the types of its members. That is, the more similar a coalition's members are (where one dimension can have greater weight than the other in computing similarity) the higher its value. Additionally, I scale the value of a three-judge coalition by α—representing the fact that unanimous opinions are more valuable than split ones—and let $d \geq 0$ (for *dissent*) be the value of a one-judge coalition.

[2]Because I focus on dissents, I ignore the role of concurrences in potentially creating a partition into three groups.

I now impose equilibrium conditions for an opinion to be stable. Intuitively, an opinion must distribute all available surplus to the judges and be immune to deviations from judges within the coalition. More formally, let C be the majority coalition in an opinion. For that opinion to be stable, I require

$$\sum_{i \in C} u_i = v(C) \qquad \text{and} \qquad u_i \geq d \, \forall \, i \in C,$$

so that all value generated by the majority coalition is shared among its members, each of whom get at least the payoff from their outside option of dissenting. To ensure a unanimous coalition is not vulnerable to a deviation by a pair of judges, I require

$$\text{If } |C| = 3 : \sum_{i \in C} u_i \geq v(D) \, \forall \, D \subseteq C.$$

Majority coalitions of three, two, or one judge(s) are all possible depending on the values of d and α. To check whether the unanimous majority coalition forms, it suffices to check whether the value $v(C)$ of that coalition and the value $v(D)$ of the coalition D of the two most similar judges satisfy $v(C) \geq v(D) + d$. While D is the key coalition to check, more than one two-judge coalition can be stable. For example, with two-dimensional judge types that have equal weights, judge types $t_1 = (0, 0)$, $t_2 = (0, 1)$ and $t_3 = (1, 0)$ produce coalitions $\{1, 2\}$ and $\{1, 3\}$ which give the same value to their members.

In Section 3.5.3, I describe two ways of equalizing dissent rates among majority and minority partisans when the weight on party is smaller than the weight on law school, but not infinitely so. The first is to add party-level noise. To show this result, consider observed party to be $x_i \, \{0, 1\}$ (representing D and R respectively) but true party to be $x_i + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2$. Then in some cases a Democratic judge will be the partisan outlier, reducing the share of dissents by Republican judges. More generally, if there is more noise in the partisan dimension than the law school dimension, this effect will still apply in expectation.

The second way of equalizing dissent rates is to introduce anti-correlation between party and law school. Then double-minority judges will become less common, and the increased dissent rate by law-school-minority judges compared to minority partisans will have a larger effect on total dissents. In the balance tables of C.2.1, I show that there is a slight anti-correlation between bipartisan panels and panels with a law school minority judge, though the correlation between bipartisan panels and panels with a single woman is slightly positive.

In the analysis of Section 3.5.3, I do not explicitly select among the potential multiplicity of stable opinions, but informally consider opinions that deliver greater total value to the judges to be more likely than ones that deliver lower total value, in the style of random utility models. Establishing an equilibrium selection rule is an important direction for future theoretical work. Another important next step is to include ideology scores like JCS and DIME-CF, e.g., by considering them as a one-dimensional embedding of a richer type space with more dimensions for possible judges characteristic.

C.4 Evaluation and Validation of Machine Learning Approaches

Below I describe a typical fine-tuning pipeline:

1. Remove names of people and locations using a standard named-entity recognizer, and manually remove the names of all judges in the sample (including those not associated with a given opinion). This step ensures that embeddings reflect the content of the text and do not "cheat" by learning and using, e.g., the author name to predict party, or the geographic location to predict political polarization.

2. Divide opinions into overlapping chunks based on the maximum length of the relevant model (in the case of transformer models, this length is usually 512 words).

3. Randomly select 25% of the opinion chunks (approximately 60,000 full opinions or 350,000 512-word chunks) to use for training. I refer to this sample as the fine-tuning sample, and the remaining 75% of the opinions as the validation sample.

4. Run a standard fine-tuning loop (for transformer models, done in Pytorch) or few-shot prediction task (for LLMs, done using the appropriate API).

5. Merge opinion chunks by averaging embeddings across all validation chunks in an opinion and re-classifying based on the new opinion-level embedding.

6. Perform additional validation (see discussion below).

7. Select a model for use based on the appropriateness of the training task and the performance in validation tasks.

Throughout this work, when using NLP measures I report results holding out opinion chunks from the fine-tuning sample. I am still able to obtain results for almost all opinions in this case (since only opinions where all chunks were used for fine-tuning are omitted) and all results are robust to including fine-tuning opinion chunks.

Model validation and selection is done in four main ways. First, for transformer models, the fine-tuning loop automatically splits the fine-tuning sample into training, testing, and validation subsets. A successful fine-tuning run produces a model that achieves good results (under a metric like accuracy or F1 score) on the validation subset of the fine-tuning sample. Second, I can evaluate those same metrics on the large validation sample. Third, I can perform permutation inference by drawing random labels for the fine-tuning sample from the empirical label distribution, fine-tuning new "permuted models" on those relabeled opinions, then evaluating the permuted models on the validation sample (with true labels). This process is extremely computationally intensive, since it requires fine-tuning hundreds or even thousands of models to produce a distribution of permuted model results. Finally, I can train a simple dense neural network (or other statistical model) to perform cross-task predictions—for example, after fine-tuning a model to predict legal topics, I can use those embeddings to predict author party. This approach does not validate model performance on the target task, but instead assesses generalizability of the learned embeddings. Below I

focus on the second approach, presenting out-of-sample performance on a variety of standard metrics using the validation sample.

Table C.15: Confusion Matrix for Predicted Legal Topic

	Predicted Label								
True Label	Crim.	Civ. Rights	1st Am.	Due Process	Priv.	Labor	Econ./Reg.	Misc.	Sum
Crim.	64,156	3,968	16	1,471	0	15	1,101	438	71,165
Civ. Rights	2,902	9,255	24	5,444	0	133	586	649	18,993
1st Am.	103	22	1,226	577	0	16	333	39	2,316
Due Process	4,196	3,615	321	41,535	0	2,805	22,828	2,126	77,426
Priv.	13	4	0	349	0	0	23	3	392
Labor	48	36	1	3,473	0	13,846	2,181	47	19,632
Econ./Reg.	1,418	437	67	8,869	0	1,289	87,041	413	99,534
Misc.	1,195	889	55	3,922	1	199	4,622	8,967	19,850
Sum	74,031	18,226	1,710	65,640	1	18,303	118,715	12,286	308,912

Results from the best-performing transformer model for topic prediction, fine-tuned Legal-BERT from Chalkidis et al. [23]. Overall, 226,026 out of 308,912 opinions (73.2%) were classified correctly.

Table C.16: Performance Metrics by Topic

True Label	Accuracy	Precision	Recall	F1 Score
Crim.	0.945	0.867	0.902	0.884
Civ. Rights	0.940	0.508	0.487	0.497
1st Am.	0.995	0.717	0.529	0.609
Due Process	0.806	0.633	0.536	0.581
Priv.	0.999	0.000	0.000	0.000
Labor	0.967	0.756	0.705	0.730
Econ./Reg.	0.857	0.733	0.874	0.798
Misc.	0.953	0.707	0.452	0.551
Weighted avg.	0.884	0.724	0.731	0.722

Performance metrics by topic for the best-performing transformer model for topic prediction, fine-tuned LegalBERT from Chalkidis et al. [23]. Of the three main topics—criminal, due process, and economics/regulation—the first and last show strong performance, and the weighted averages reflects an overall successful classifier. Weaker performance on due process cases is unsurprising since that topic often overlaps with others, and reflect realistically higher weights for other topics in predictions for due process cases.

C.5 Robustness Checks for Dissents

In this appendix, I provide robustness checks for results about dissent rates and dissent shares. In Appendix C.5.1, I separate out results by Democratic and Republican judges, replicating tables C.2 (Section 3.6.1) and C.3 (Section 3.6.2) from the main text. In Appendix C.5.2, I estimate judge-level dissents as in Table C.4 (Section 3.6.3) using alternative specifications; I also plot the amount of minority-party and median dissents over time. In Appendix C.5.3, I provide additional analysis comparing outlier effects in the style of Table C.7.

C.5.1 Separating Partisan Dissents for Democrats and Republicans

Table C.17: Replicating Table C.2 with Democratic-Majority Panels

	(1) Party	(2)	(3) JCS Scores	(4)	(5)	(6) DIME-CF Scores	(7)
	All Panels	All Panels	Party-Unan.	Bip.	All Panels	Party-Unan.	Bip.
bipartisan	0.128^{***}						
	(0.031)						
ext. dist.		0.538^{***}	0.494^{**}	0.296^{**}	0.147^{***}	0.018	0.062
		(0.071)	(0.246)	(0.117)	(0.044)	(0.111)	(0.056)
2nd-ext. dist.		0.165	1.582^{***}	−0.075	0.078	0.154	0.029
		(0.212)	(0.609)	(0.232)	(0.088)	(0.297)	(0.090)
p for coeff. diff.	–	0.121	0.130	0.178	0.489	0.697	0.762
Control dissents	0.099	0.093	0.088	0.119	0.097	0.104	0.100
Score std. dev.	–	0.359	0.341	0.353	0.663	0.628	0.655
Observations	111, 142	45, 339	8, 660	35, 601	47, 288	9, 348	36, 890

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

The effects of Table C.2 persist when considering Democratic-majority panels only. All estimation is performed as in that table with an appropriate modification of the sample.

Table C.18: Replicating Table C.2 with Republican-Majority Panels

	(1) Party	(2)	(3) JCS Scores	(4)	(5)	(6) DIME-CF Scores	(7)
	All Panels	All Panels	Party-Unan.	Bip.	All Panels	Party-Unan.	Bip.
bipartisan	0.323***						
	(0.029)						
ext. dist.		0.722***	−0.090	0.411***	0.261***	0.016	0.032
		(0.069)	(0.188)	(0.124)	(0.031)	(0.075)	(0.048)
2nd-ext. dist.		0.094	−0.334	0.152	0.186**	−0.175	0.229***
		(0.113)	(0.313)	(0.129)	(0.076)	(0.192)	(0.087)
p for coeff. diff.	–	0.000	0.547	0.156	0.399	0.393	0.071
Control dissents	0.075	0.072	0.071	0.106	0.073	0.064	0.099
Score std. dev.	–	0.355	0.355	0.354	0.671	0.659	0.665
Observations	143, 450	82, 951	29, 183	52, 502	84, 741	29, 844	53, 702

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

The effects of Table C.2 persist when considering Republican-majority panels only. All estimation is performed as in that table with an appropriate modification of the sample.

Table C.19: Replicating Table C.3 with Democratic-Majority Panels

		(1)	(2) JCS Scores	(3)	(4)	(5)	(6) DIME-CF Scores	(7)	(8)
Outcome	Variable	All Panels	All Panels	Party-Unan.	Bip.	All Panels	All Panels	Party-Unan.	Bip.
dissent by median judge	bipartisan	0.193***				0.270***			
		(0.064)				(0.062)			
	ext. dist.		0.419***	0.612	0.315**		0.130**	0.117	0.027
			(0.091)	(0.404)	(0.137)		(0.057)	(0.147)	(0.074)
	2nd-ext. dist. med		0.287	1.638*	0.138		0.109	0.016	0.063
			(0.306)	(0.879)	(0.332)		(0.119)	(0.393)	(0.128)
dissent by 2nd-extremal judge	bipartisan	0.251***				0.188***			
		(0.062)				(0.061)			
	ext. dist.		0.449***	0.704**	0.201		0.084	−0.122	0.018
			(0.087)	(0.314)	(0.134)		(0.057)	(0.166)	(0.072)
	2nd-ext. dist.		0.075	1.288	−0.194		0.036	0.285	−0.010
			(0.289)	(0.881)	(0.317)		(0.117)	(0.372)	(0.126)
dissent by extremal judge	bipartisan	0.350***				0.240***			
		(0.066)				(0.061)			
	ext. dist.		0.500***	−0.187	0.238*		0.158***	0.011	0.112
			(0.089)	(0.409)	(0.133)		(0.058)	(0.154)	(0.074)
	2nd-ext. dist.		0.140	1.083	−0.057		0.042	0.136	0.003
			(0.291)	(1.088)	(0.312)		(0.120)	(0.365)	(0.128)
	bip./ext. = across roles	0.238	0.819	0.213	0.841	0.644	0.668	0.574	0.623
	2nd-ext. = across roles	–	0.881	0.922	0.778	–	0.894	0.888	0.915
	Control dissents	0.032	0.090	0.089	0.096	0.034	0.096	0.100	0.090
	Score std. dev.	–	0.359	0.195	0.380	–	0.662	0.574	0.678
	Observations	107, 498	46, 187	9, 775	36, 412	107, 684	48, 149	10, 464	37, 685

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Replication of Table C.3 restricting to Democratic-majority panels only. All estimation is performed as in that table with an appropriate modification of the sample. In column (1), using JCS scores, the increase in dissents on bipartisan panels compared to party-unanimous panels is largest for extremal judges and smallest for median judges. However, this difference is not in evidence for any other columns, and is only significant at the 10% level when comparing the extremal and median judges.

Table C.20: Replicating Table C.3 with Republican-Majority Panels

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		JCS Scores				DIME-CF Scores			
Outcome	Variable	All Panels	All Panels	Party-Unan.	Bip.	All Panels	All Panels	Party-Unan.	Bip.
dissent by median judge	bipartisan	0.392***				0.336***			
		(0.043)				(0.043)			
	ext. dist.		0.683***	0.235	0.312**		0.262***	0.105	0.047
			(0.086)	(0.270)	(0.158)		(0.041)	(0.111)	(0.059)
	2nd-ext. dist.		0.077	−0.613	0.073		0.129	−0.137	0.175
			(0.157)	(0.500)	(0.177)		(0.098)	(0.247)	(0.111)
dissent by 2nd-extremal judge	bipartisan	0.378***				0.392***			
		(0.043)				(0.043)			
	ext. dist.		0.630***	−0.386	0.356**		0.243***	−0.084	0.051
			(0.089)	(0.291)	(0.171)		(0.042)	(0.125)	(0.060)
	2nd-ext. dist.		0.062	−0.027	0.056		0.222**	0.260	0.224**
			(0.160)	(0.508)	(0.180)		(0.100)	(0.263)	(0.111)
dissent by extremal judge	bipartisan	0.296***				0.342***			
		(0.043)				(0.043)			
	ext. dist.		0.587***	−0.074	0.390**		0.181***	0.029	−0.028
			(0.088)	(0.253)	(0.164)		(0.042)	(0.113)	(0.060)
	2nd-ext. dist.		0.108	−0.343	0.297*		0.159	−0.611**	0.224**
			(0.155)	(0.452)	(0.177)		(0.098)	(0.266)	(0.109)
	bip./ext. = across roles	0.238	0.746	0.303	0.944	0.601	0.367	0.537	0.586
	2nd-ext. = across roles	–	0.979	0.720	0.579	–	0.806	0.072	0.939
	Control dissents	0.048	0.071	0.069	0.095	0.049	0.072	0.064	0.099
	Score std. dev.	–	0.355	0.240	0.387	–	0.671	0.543	0.703
	Observations	141, 195	83, 379	30, 496	52, 883	141, 359	85, 149	31, 066	54, 083

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

The effects of Table C.3 persist when considering Republican-majority panels only. All estimation is performed as in that table with an appropriate modification of the sample.

Table C.21: Replicating Table C.3 with Separate Democratic-Majority and Republican-Majority Bipartisan Panels

		(1) JCS	(2) DIME-CF
dissent by median judge	Rep-maj.	0.313***	0.285***
		(0.038)	(0.037)
	Dem-maj.	0.333***	0.350***
		(0.041)	(0.041)
dissent by 2nd-extremal judge	Rep-maj.	0.288***	0.295***
		(0.038)	(0.037)
	Dem-maj.	0.403***	0.364***
		(0.041)	(0.041)
dissent by extremal judge	Rep-maj.	0.268***	0.268***
		(0.038)	(0.037)
	Dem-maj.	0.380***	0.369***
		(0.041)	(0.040)
	Control dissents	0.042	0.044
	Observations	248,693	249,043

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Columns (1) and (2) replicate columns (1) and (5), respectively, of Table C.3 with two separate indicators for Democratic-majority and Republican-majority bipartisan panels. The estimation method and sample are the same as in those columns. I perform a Wald test for equality of the three coefficients within each of the four score × party categories, then for corresponding Democratic and Republican coefficients across all six judge role × score pairs. These results show that dissent rates are equal across judge roles within party, but differ within judge role across parties.

	JCS	DIME-CF
Rep.	0.710	0.879
Dem.	0.484	0.944

	JCS	DIME-CF
median	0.572	0.001
2nd-extremal	0.001	0.059
extremal	0.047	0.004

C.5.2 Dissents at the Judge Level

Table C.22: Judge-Level Dissents with Controls

	(1) Min. Party	(2) Is JCS Median	(3) Is DIME-CF Median	(4) Min. Law School	(5) Min. Gender	(6) Only Man	(7) Only Woman	(8) Youngest	(9) Median Age	(10) Oldest
judge char.	0.001	−0.002	−0.005	−0.005	−0.005	−0.036	0.145***	0.009	−0.009	−0.001
	(0.017)	(0.019)	(0.019)	(0.029)	(0.044)	(0.074)	(0.038)	(0.014)	(0.014)	(0.013)
is Dem.	0.008	−0.018	−0.018	−0.016	−0.022	−0.022	−0.021	−0.008	−0.008	−0.008
	(0.016)	(0.022)	(0.021)	(0.034)	(0.034)	(0.034)	(0.034)	(0.016)	(0.016)	(0.016)
age	−0.027**	−0.068***	−0.064***	−0.005	−0.034	−0.034	−0.033	_	_	_
	(0.012)	(0.019)	(0.019)	(0.026)	(0.022)	(0.022)	(0.022)	_	_	_
age^2	0.000**	0.001***	0.001***	0.000	0.000	0.000	0.000	_	_	_
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)			
is woman	0.165***	0.147***	0.147***	0.128	0.181***	_	_	0.171***	0.171***	0.171***
	(0.044)	(0.042)	(0.042)	(0.083)	(0.050)	_	_	(0.040)	(0.040)	(0.040)
Control dissents	0.031	0.035	0.035	0.031	0.030	0.031	0.030	0.032	0.032	0.032
Observations	533, 214	389, 766	398, 745	184, 767	144, 555	144, 555	144, 555	772, 230	772, 230	772, 230

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Replicating Table C.4 with controls for judge characteristics instead of judge fixed effects. Controls for panel characteristics—number of Democrats, number of unique law schools, number of women, and Bloomberg topic, as appropriate—are also included.

Table C.23: Judge-Level Dissents with Fixed Effects, Bipartisan Panels

	(1) Min. Party	(2) Is JCS Median	(3) Is DIME-CF Median	(4) Min. Law School	(5) Min. Gender	(6) Only Man	(7) Only Woman	(8) Youngest	(9) Median Age	(10) Oldest
judge char.	0.010	0.019	0.002	0.009	0.122^{***}	-0.013	0.136^{***}	-0.024	0.009	0.011
	(0.022)	(0.026)	(0.026)	(0.050)	(0.046)	(0.084)	(0.050)	(0.025)	(0.020)	(0.026)
Control dissents	0.035	0.039	0.039	0.034	0.033	0.033	0.033	0.034	0.035	0.035
Observations	533, 214	389, 766	398, 745	184, 767	144, 555	144, 555	144, 555	772, 230	772, 230	772, 230
# judges	1, 230	756	771	710	592	592	592	1, 230	1, 230	1, 230

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Replicating Table C.4, restricted to bipartisan panels only. Controls for panel characteristics—number of Democrats, number of unique law schools, number of women, and Bloomberg topic, as appropriate—are also included.

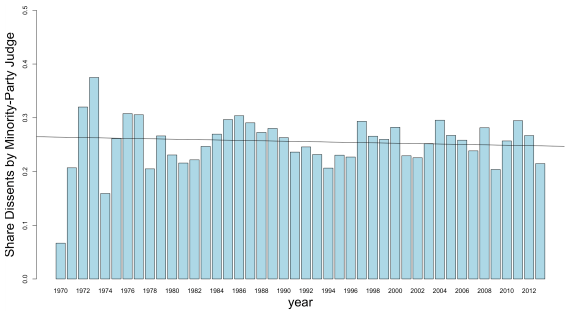
Table C.24: Judge-Level Dissents with Controls, Bipartisan Panels

	(1) Min. Party	(2) Is JCS Median	(3) Is DIME-CF Median	(4) Min. Law School	(5) Min. Gender	(6) Only Man	(7) Only Woman	(8) Youngest	(9) Median Age	(10) Oldest
judge char.	0.001	-0.010	-0.005	0.014	-0.007	-0.004	0.135^{***}	0.006	-0.009	0.003
	(0.017)	(0.021)	(0.021)	(0.035)	(0.051)	(0.080)	(0.044)	(0.016)	(0.016)	(0.016)
is Dem.	0.008	-0.017	-0.017	-0.017	-0.022	-0.022	-0.021	-0.008	-0.008	-0.008
	(0.016)	(0.022)	(0.022)	(0.034)	(0.034)	(0.034)	(0.034)	(0.016)	(0.016)	(0.016)
age	-0.027^{**}	-0.052^{**}	-0.047^{**}	0.002	-0.034	-0.035	-0.034	_	_	_
	(0.012)	(0.021)	(0.021)	(0.032)	(0.025)	(0.025)	(0.025)	_	_	_
age^2	0.000^{**}	0.000^{**}	0.000^{**}	0.000	0.000	0.000	0.000	_	_	_
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)			
is woman	0.165^{***}	0.122^{***}	0.122^{***}	0.194^{**}	0.180^{***}	_	_	0.165^{***}	0.165^{***}	0.165^{***}
	(0.044)	(0.046)	(0.046)	(0.093)	(0.054)	_	_	(0.044)	(0.044)	(0.044)
Control dissents	0.034	0.038	0.038	0.034	0.033	0.033	0.033	0.034	0.034	0.034
Observations	533, 214	267, 420	274, 416	125, 121	105, 492	105, 492	105, 492	533, 214	533, 214	533, 214

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Replicating Table C.4 with controls for judge characteristics instead of judge fixed effects, and restricted to bipartisan panels only. Controls for panel characteristics—number of Democrats, number of unique law schools, number of women, and Bloomberg topic, as appropriate—are also included.

Figure C.18: Minority-Party Dissents Over Time

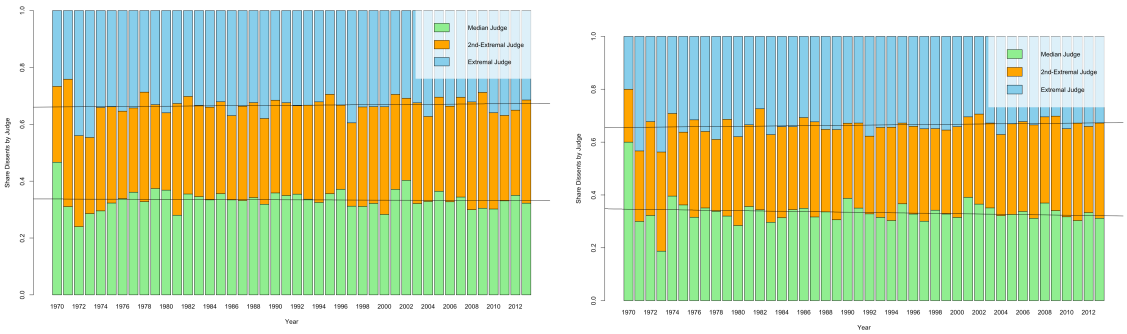


Share of dissents on bipartisan panels authored by the minority-party judge, 1970–2013. The line is a linear regression of minority-party dissent share on year,

$$\text{dissent} = 0.264 - 0.0003\,\text{yr}.$$

The slope is not significantly different from 0 at the 10% level.

Figure C.19: Dissents by Score Role Over Time



Share of dissents on bipartisan panels authored by each score role—median, second-extremal, extremal—using JCS scores (left) and DIME-CF scores (right), 1970–2013. The lines are linear regressions of dissent share on year. For JCS scores (left),

$$\text{dissent} = 0.339 - 0.0002\,\text{yr} \qquad \text{and} \qquad \text{dissent} = 0.338 - 0.0001\,\text{yr},$$

for the extremal and median judges, respectively. For DIME-CF scores (right),

$$\text{dissent} = 0.344 - 0.0003\,\text{yr} \qquad \text{and} \qquad \text{dissent} = 0.347 - 0.0004\,\text{yr},$$

for the extremal and median judges, respectively. No slope is significantly different from 0 the 10% level.

C.5.3 Comparing Different Outlier Effects

Table C.25: Comparison of Outlier Effects Interacted with Party

	(1) Partisanship	(2) Shared Law School	(3) Any Law Schools	(4) Mixed Gender	(5) Min. Men	(6) Min. Women	(7) Age
bipartisan	0.232^{***}						
	(0.023)						
one diff. law school, panel-maj. party		0.078	-0.092^{***}				
		(0.080)	(0.031)				
one diff. law school panel-min. party		0.200^{**}	0.036				
		(0.084)	(0.036)				
all same law school			-0.177^{**}				
			(0.079)				
panel-min. gender, panel-maj. party				0.051			
				(0.032)			
panel-min. gender, panel-min. party				0.036			
				(0.039)			
one man, panel-maj. party					0.146		
					(0.181)		
one man, panel-min. party					0.098		
					(0.199)		
one woman, panel-maj. party						0.032	
						(0.033)	
one woman, panel-min. party						0.024	
						(0.040)	
young outlier, panel-maj. party							-0.010
							(0.041)
young outlier, panel-min. party							0.099^{*}
							(0.051)
old outlier, panel-maj. party							0.000
							(0.042)
old outlier, panel-min. party							0.113^{*}
							(0.058)
p for coeff. diff.	–	0.012	0.004	0.695	0.604	0.854	0.432 (young); 0.159 (old)
Control dissents	0.084	0.067	0.107	0.098	0.161	0.098	0.102
Observations	178,037	48,720	178,037	178,037	3,848	173,426	178,037

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Effect of panel composition on total dissents, considering bipartisan panels only and splitting each explanatory variable by whether the outlier judge is a majority or minority partisan. Column (1) shows the effect of partisanship for reference. Columns (2)–(6) use the subset of corresponding panels from Table C.7 which are bipartisan. All columns follow the same estimation method as in that table.

To directly compare effect sizes (as much as is possible), I estimate a conditional logit model including all coefficients together:

$$\begin{aligned}\text{Dissent}_p = &\beta_1 \text{bipartisan}_p \\ &+ \beta_2 3 \text{ same law school}_p + \beta_3 2 \text{ same law school}_p + \\ &\beta_4 \text{young outlier} + \beta_5 \text{old outlier}_p \\ &+ \beta_6 \text{one man}_p + \beta_7 \text{one woman}_p \\ &+ \text{legal-topic}_p + \gamma_{ct} + \varepsilon_{pct}.\end{aligned} \tag{C.1}$$

Results are in Table C.26.

Table C.26: Direct Comparison of Outlier Effects

Bipartisan	One Unique Law School.	2 Unique Law Schools	One Man	One Woman	Young Outlier	Old Outlier
0.233^{***}	-0.087	-0.027	0.202^{***}	0.080^{**}	0.042	0.044
(0.023)	(0.066)	(0.021)	(0.052)	(0.025)	(0.030)	(0.030)

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

There are 257,712 panels in the sample, and the baseline dissent rate is 8.2%. Party and gender have the strongest effects, with the other coefficients not significant at the 10% level. Averaging across categories for a back-of-the-envelope calculation, shared law school training reduces dissents by 5.1%, a mixed-gender panel increases dissents by 13.8%, and an age outlier increases dissents by 4.0%. These effects sum to 23.1%, comparable to the 23.5% for party but with a much wider confidence interval.

Table C.27: Dissents and Deference to Lower-Court Rulings

		Party	JCS Scores			DIME-CF Scores		
Outcome	Variable	All Panels	All Panels	Party-Unan.	Bip.	All Panels	Party-Unan.	Bip.
uphold w/ dissent	bipartisan	0.254^{***}						
		(0.021)						
	ext. dist.		0.537^{***}	0.070	0.195^{**}	0.208^{***}	0.014	0.041
			(0.045)	(0.168)	(0.079)	(0.025)	(0.072)	(0.036)
	2nd-ext. dist.		0.270^{***}	-0.175	0.252^{**}	0.224^{***}	-0.167	0.195^{***}
			(0.094)	(0.341)	(0.112)	(0.056)	(0.174)	(0.065)
reverse w/ dissent	bipartisan	0.153^{***}						
		(0.021)						
	ext. dist.		0.536^{***}	0.103	0.462^{***}	0.186^{***}	0.016	0.072^{*}
			(0.046)	(0.170)	(0.080)	(0.026)	(0.072)	(0.037)
	2nd-ext. dist.		-0.119	-0.064	-0.165	0.070	-0.038	0.025
			(0.097)	(0.328)	(0.117)	(0.057)	(0.160)	(0.067)
reverse w/ dissent	bipartisan	-0.027^{***}						
		(0.007)						
	ext. dist.		-0.076^{***}	-0.218^{***}	-0.012	-0.025^{***}	0.022	-0.019
			(0.015)	(0.050)	(0.029)	(0.009)	(0.021)	(0.013)
	2nd-ext. dist.		-0.155^{***}	-0.156	-0.126^{***}	-0.018	0.013	-0.005
			(0.032)	(0.099)	(0.041)	(0.018)	(0.046)	(0.023)
	bip./ext. = across dis.	0.000	0.000	0.115	0.020	0.000	0.926	0.139
	2nd-ext. = across dis.	–	0.000	0.959	0.003	0.000	0.337	0.006
	Control dissents	0.083	0.077	0.076	0.097	0.078	0.073	0.095
	Score std. dev.	–	0.388	0.351	0.398	0.666	0.709	0.709
	Observations	260, 123	151, 082	40, 299	89, 417	155, 378	41, 561	91, 891

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Results from conditional multinomial logit estimation of dissent rates using circuit × year fixed effects (with robust standard errors clustered at the circuit × year level), in the style of equations (3.4) and (3.5) but with the four categories representing uphold or reverse × unanimously or with dissent. Increased dissents on bipartisan panels come mostly from upholdings; this disparity vanishes or even reverses when looking at the distance between the extremal and median judges.

C.6 Robustness Checks for NLP Results

In this appendix, I provide robustness checks for the results from Section 3.8. In Appendix C.6.1, I show that the differences in uniform and Bayesian major topic use are robust to scaling those thresholds. and provide distributions of major topics using the un-scaled measures. In Appendix C.6.2, I replicate Table C.10 (Section 3.8.1) with author controls and fixed effects. Finally, in Appendix C.6.3, I provide distributions of neighbor match scores from Table C.11 (Section 3.8.2), and in Appendix C.6.4 I show that the results of that table are robust to alternative specifications of the nearest-neighbor measure.

C.6.1 Robustness of Major Topic Measures

Table C.28: Robustness of Uniform Major Topics to Alternative Thresholds

Metric	All-D. vs. All-R.	2D. vs. 3D.	2R. vs. 3R.	Bip. D. vs. R.	Sample Mean
# topics w/ $\mathbb{P} > 0.25$	-0.001	0.003	0.005*	0.001	1.388
	(0.006)	(0.004)	(0.003)	(0.003)	
# topics w/ $\mathbb{P} > 0.125$	0.001	0.006	0.005*	0.003	1.772
	(0.006)	(0.004)	(0.003)	(0.003)	
# topics w/ $\mathbb{P} > 0.1$	0.003	0.005	0.007**	0.003	1.906
	(0.006)	(0.004)	(0.003)	(0.003)	
# topics w/ $\mathbb{P} > 0.05$	0.013**	-0.006	0.006**	0.004	2.281
	(0.007)	(0.004)	(0.003)	(0.003)	
# topics w/ $\mathbb{P} > 0.01$	0.015**	-0.007*	0.006*	0.004	3.250
	(0.005)	(0.004)	(0.003)	(0.002)	
Observations	49,099	61,538	101,862	114,301	189,538

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Robustness to alternative uniform thresholds for major topics. A threshold of 0.125 (reported in the main text, and again here) represents a uniform distribution over all 8 topics. The number of major topics on all-Democratic panels compared to all-Republican panels grows slightly as the threshold decreases, but the difference remains small; the number of major topics is the same on bipartisan and party-unanimous panels.

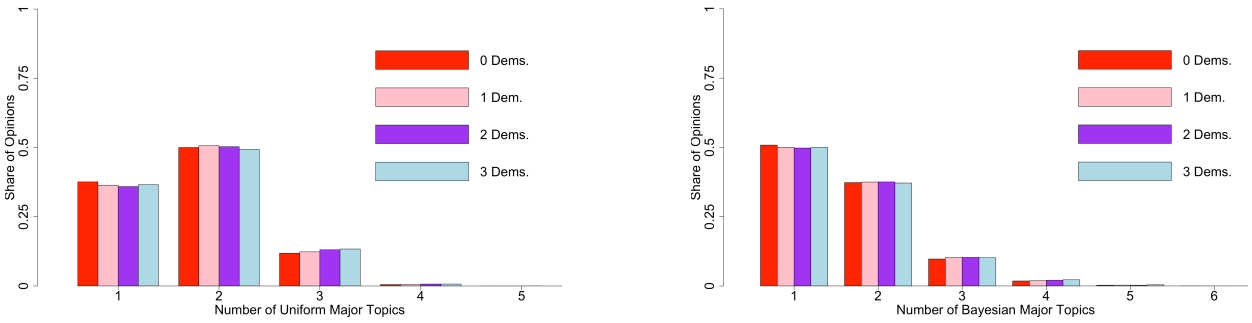
Table C.29: Robustness of Bayesian Major Topics to Alternative Thresholds

Metric	All-D. vs. All-R.	2D. vs. 3D.	2R. vs. 3R.	Bip. D. vs. R.	Sample Mean
# topics w/ posterior > prior	0.014*	-0.007	0.007*	0.004	1.651
	(0.007)	(0.005)	(0.004)	(0.003)	
# topics w/ posterior > 1.25×prior	0.014**	-0.002	0.007**	0.004	1.480
	(0.007)	(0.005)	(0.003)	(0.003)	
# topics w/ posterior > 1.5×prior	0.006	0.001	0.005	0.003	1.331
	(0.006)	(0.005)	(0.003)	(0.003)	
# topics w/ posterior > 2×prior	0.014**	-0.008**	0.000	0.000	1.010
	(0.005)	(0.004)	(0.003)	(0.002)	
# topics w/ posterior > 3×prior	0.025**	-0.014*	-0.001	-0.005	0.735
	(0.010)	(0.008)	(0.005)	(0.005)	
Observations	49, 099	61, 535	101, 862	114, 301	189,538

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

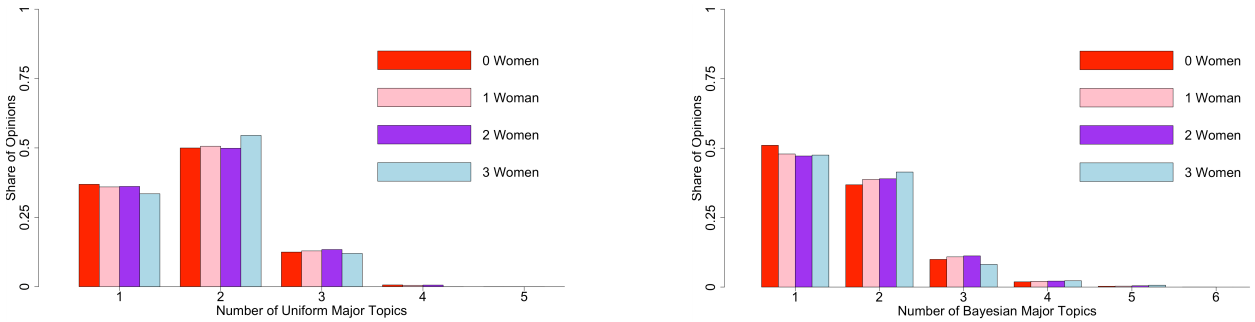
Robustness to alternative Bayesian thresholds (posterior $> \alpha\times$ prior) for major topics. A value of $\alpha = 1$ (reported in the main text, and again here) represents those topics where the posterior distribution places more weight than the prior. Across thresholds, all-Democratic panels have slightly more major topics than all-Republican panels, and bipartisan panels have the same number of topics as party-unanimous panels.

Figure C.20: Comparing Number of Major Topics By Panel Party Composition



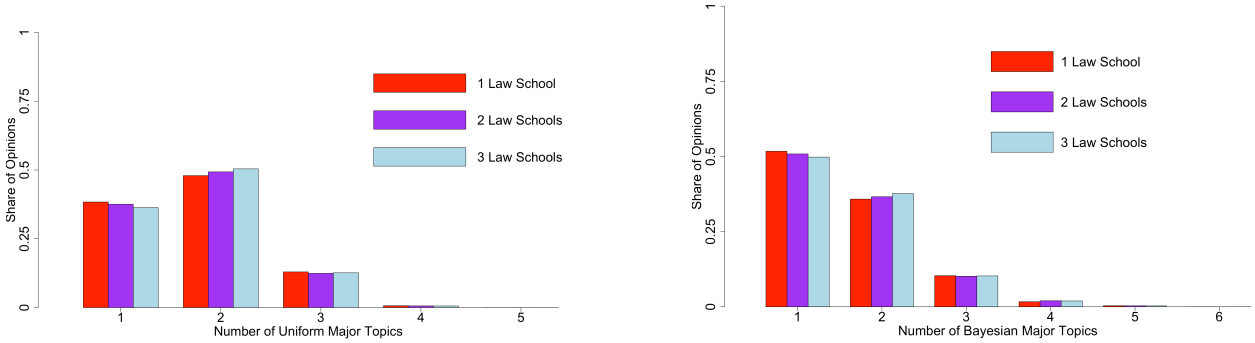
In the left image, distributions of topics with greater-than-uniform probability across panels, 1970–2013 (sample of 163,400) by number of Democrats (means 1.753, 1.771, 1.786, and 1.782 respectively; median 2 for all panel types). In the right image, distributions of topics with greater-than-prior probability across panels, same sample (means 1.635, 1.649, 1.656, and 1.660 respectively; medians 1, 1, 2, and 2 respectively).

Figure C.21: Comparing Number of Major Topics By Panel Gender Composition



In the left image, distributions of topics with greater-than-uniform probability across panels, 1970–2013 (sample of 188,551) by number of women (means 1.775, 1.761, 1.761, and 1.779 respectively; median 2 for all panel types). In the right image, distributions of topics with greater-than-prior probability across panels, same sample (means 1.651, 1.649, 1.666, and 1.713 respectively; medians 1, 1, 2, 2 respectively).

Figure C.22: Comparing Number of Major Topics By Panel Law School Composition



In the left image, distributions of topics with greater-than-uniform probability across panels, 1970–2013 (sample of 188,551) by number of unique law schools represented (means 1.631, 1.644, and 1.653 respectively' medians 1, 2, and 2 respectively). In the right image, distributions of topics with greater-than-prior probability across panels, same sample (means 1.631, 1.644 and 1.653 respectively; medians 1, 2, and 2 respectively).

C.6.2 Robustness of Topic Distributions from Section 3.8.1 to Alternative Controls

Table C.30: Comparing Topics Across All-Democratic and All-Republican Panels, Author Controls

	(1) Crim.	(2) Civ Rights	(3) 1A	(4) Due Process	(5) Priv.	(6) Lab. Rel.	(7) Econ. & Reg.	(8) Misc.	(9) Unif.	(10) Bayes
is Dem.	−0.013	0.034	0.252**	−0.005	−0.007	0.030	−0.065**	0.163***	0.002	0.015***
	(0.035)	(0.034)	(0.120)	(0.020)	(0.165)	(0.048)	(0.025)	(0.049)	(0.006)	(0.045)
age	0.022	0.003	−0.028	−0.013*	0.119	−0.001	−0.006	0.022	−0.003	0.020
	(0.016)	(0.018)	(0.069)	(0.008)	(0.094)	(0.023)	(0.012)	(0.022)	(0.003)	(0.021)
age^2	−0.000	−0.000	0.000	0.000*	−0.001	0.000	0.000	−0.000	0.000	−0.000
	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
is woman	0.096**	0.165***	0.033	0.007	0.001	−0.164**	−0.125***	−0.015	0.009	−0.014
	(0.044)	(0.044)	(0.149)	(0.024)	(0.177)	(0.071)	(0.036)	(0.057)	(0.008)	(0.052)
Control mean	0.278	0.070	0.006	0.293	0.001	0.061	0.241	0.050	1.754	1.631

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Replicating the first row of Table C.10 with controls for author characteristics; estimation follows that table.

Table C.31: Comparing Topics Across Democratic-Majority Panels, Author Controls

	(1) Crim.	(2) Civ Rights	(3) 1A	(4) Due Process	(5) Priv.	(6) Lab. Rel.	(7) Econ. & Reg.	(8) Misc.	(9) Unif.	(10) Bayes
bipartisan	0.009	−0.021	−0.119	−0.000	−0.070	−0.023	0.043**	−0.098***	0.005	−0.008
	(0.024)	(0.025)	(0.075)	(0.014)	(0.114)	(0.035)	(0.019)	(0.036)	(0.004)	(0.005)
is Dem.	0.023	−0.038*	−0.015	−0.010	−0.176*	0.018	0.001	−0.008	−0.001	−0.006
	(0.019)	(0.021)	(0.071)	(0.012)	(0.094)	(0.029)	(0.016)	(0.028)	(0.004)	(0.004)
age	0.006	−0.006	−0.045	0.008	0.092	0.004	−0.012	0.007	0.001	−0.002
	(0.016)	(0.016)	(0.066)	(0.007)	(0.069)	(0.021)	(0.010)	(0.019)	(0.003)	(0.003)
age^2	−0.000	0.000	0.000	−0.000	−0.001	−0.000	0.000	−0.000	−0.000	0.000
	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
is woman	−0.093***	0.079**	0.203**	0.061***	0.119	−0.042	−0.063**	0.127***	0.021***	0.019**
	(0.035)	(0.034)	(0.097)	(0.021)	(0.133)	(0.053)	(0.031)	(0.043)	(0.007)	(0.008)
Control mean	0.278	0.071	0.006	0.293	0.001	0.060	0.242	0.050	1.754	1.636

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Replicating the second row of Table C.10 with controls for author characteristics; estimation follows that table.

Table C.32: Comparing Topics Across Republican-Majority Panels, Author Controls

	(1) Crim.	(2) Civ Rights	(3) 1A	(4) Due Process	(5) Priv.	(6) Lab. Rel.	(7) Econ. & Reg.	(8) Misc.	(9) Unif.	(10) Bayes
bipartisan	−0.027	0.019	0.029	0.014	0.033	−0.002	−0.000	0.020	0.005*	0.007*
	(0.017)	(0.017)	(0.063)	(0.010)	(0.084)	(0.022)	(0.012)	(0.026)	(0.003)	(0.004)
is Dem.	0.027*	−0.035*	−0.008	−0.036***	0.029	0.043*	−0.000	0.045*	−0.003	−0.004
	(0.016)	(0.018)	(0.067)	(0.009)	(0.084)	(0.026)	(0.014)	(0.024)	(0.003)	(0.003)
age	0.017	−0.002	0.085*	−0.002	−0.016	−0.007	−0.010	0.004	−0.001	0.000
	(0.011)	(0.011)	(0.051)	(0.005)	(0.048)	(0.017)	(0.007)	(0.014)	(0.002)	(0.002)
age^2	−0.000	−0.000	−0.001*	0.000	0.000	0.000	0.000	−0.000	0.000	−0.000
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
is woman	0.007	0.060**	−0.135	0.022	0.071	−0.119***	−0.033	0.039	−0.001	0.003
	(0.030)	(0.029)	(0.101)	(0.017)	(0.111)	(0.043)	(0.023)	(0.046)	(0.005)	(0.006)
Control mean	0.279	0.071	0.006	0.295	0.001	0.059	0.238	0.051	1.756	1.636

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Replicating the third row of Table C.10 with controls for author characteristics; estimation follows that table.

Table C.33: Comparing Topics Across Bipartisan Panels, Author Controls

	(1) Crim.	(2) Civ Rights	(3) 1A	(4) Due Process	(5) Priv.	(6) Lab. Rel.	(7) Econ. & Reg.	(8) Misc.	(9) Unif.	(10) Bayes
Dem. majority	−0.009	0.019	0.085*	0.013	0.030	−0.021	−0.014	0.015	0.004	0.014
	(0.017)	(0.016)	(0.052)	(0.009)	(0.074)	(0.021)	(0.013)	(0.022)	(0.003)	(0.020)
is Dem.	0.026**	−0.034**	−0.020	−0.025***	−0.063	0.032	−0.001	0.023	−0.002	0.022
	(0.012)	(0.014)	(0.050)	(0.007)	(0.066)	(0.020)	(0.011)	(0.018)	(0.002)	(0.016)
age	0.011	−0.010	0.051	0.007	−0.014	−0.005	−0.012*	0.001	0.000	0.001
	(0.011)	(0.010)	(0.046)	(0.006)	(0.043)	(0.014)	(0.007)	(0.014)	(0.002)	(0.013)
age^2	−0.000	0.000	−0.001	−0.000	0.000	0.000	0.000*	0.000	−0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
is woman	−0.073***	0.034	0.025	0.046***	0.115	−0.071*	−0.017	0.111***	0.007	0.101***
	(0.027)	(0.023)	(0.087)	(0.015)	(0.103)	(0.038)	(0.022)	(0.039)	(0.005)	(0.036)
Control mean	0.272	0.071	0.007	0.295	0.001	0.060	0.242	0.0523	1.763	1.640

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Replicating the fourth row of Table C.10 with controls for author characteristics; estimation follows that table.

Table C.34: Comparing Topic Use Across Panel Types, Author Fixed Effects

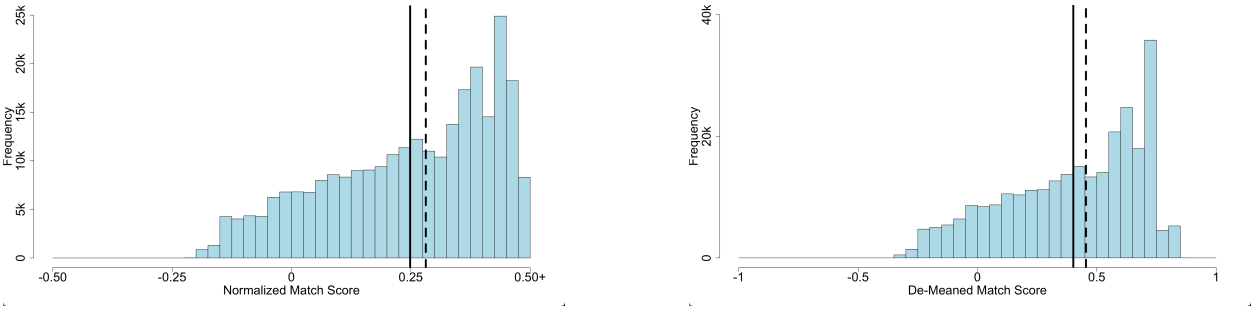
Panel type	Variable	(1) Crim.	(2) Civ Rights	(3) 1A	(4) Due Process	(5) Priv.	(6) Lab. Rel.	(7) Econ. & Reg.	(8) Misc.	(9) Unif.	(10) Bayes	Observations
majority-Dem.	2-1 Dem.	0.019	−0.016	−0.133*	−0.002	−0.106	−0.023	0.039*	−0.109***	0.003	−0.011**	
		(0.026)	(0.025)	(0.076)	(0.014)	(0.122)	(0.037)	(0.020)	(0.033)	(0.004)	(0.005)	61,494
	Control mean	0.264	0.072	0.008	0.293	0.001	0.062	0.245	0.056	1.774	1.652	
majority-Rep.	2-1 Rep.	−0.026	0.023	−0.005	0.013	0.038	−0.005	0.001	0.016	0.005*	0.007*	
		(0.016)	(0.017)	(0.067)	(0.010)	(0.082)	(0.024)	(0.014)	(0.023)	(0.003)	(0.004)	101,773
	Control mean	0.265	0.073	0.008	0.295	0.001	0.061	0.240	0.057	1.773	1.652	
bipartisan	Dem. majority	−0.008	0.025*	0.090	0.013	0.032	−0.028	−0.014	0.016	0.004	0.015	
		(0.015)	(0.015)	(0.057)	(0.009)	(0.070)	(0.024)	(0.011)	(0.020)	(0.003)	(0.018)	114,219
	Control mean	0.265	0.072	0.008	0.295	0.001	0.061	0.241	0.057	1.774	1.654	

Significance is denoted by *** for 1%, ** for 5%, and * for 10%.

Replicating rows 2–4 of Table C.10 with author fixed effects; estimation follows that table. Row 1 cannot be replicated with author fixed effects because no judge is an author on both all-Democratic and all-Republican panels.

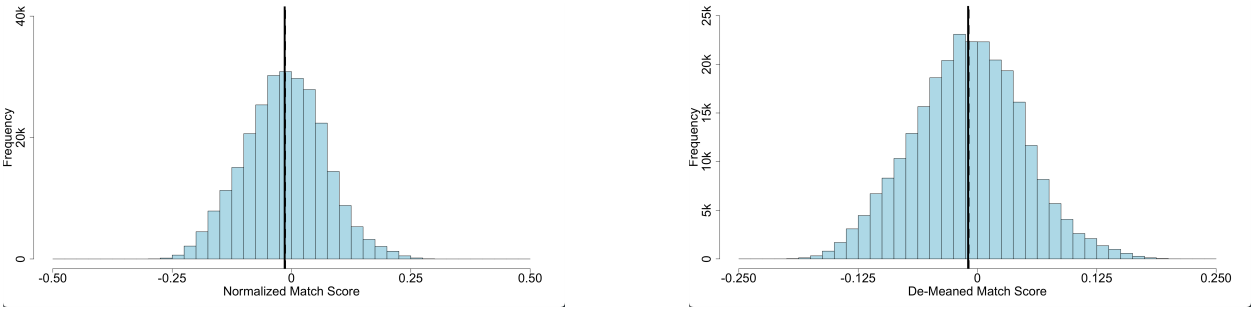
C.6.3 Distributions of Match Scores from Table C.11

Figure C.23: Distributions of Category Match Scores



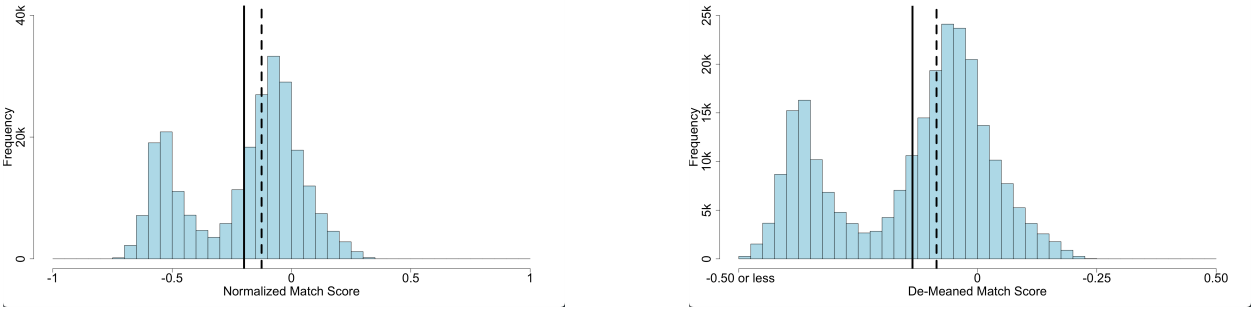
In the left image, the distribution of normalized category match scores; in the right image, the distribution of de-meaned category match scores (without dividing by the entropy of the category distribution); the sample contains 270,451 opinions. Solid lines indicate the mean of each distribution (0.249 and 0.401 respectively) and dashed lines indicate the medians (0.281 and 0.454 respectively).

Figure C.24: Distributions of Author Party Match Scores



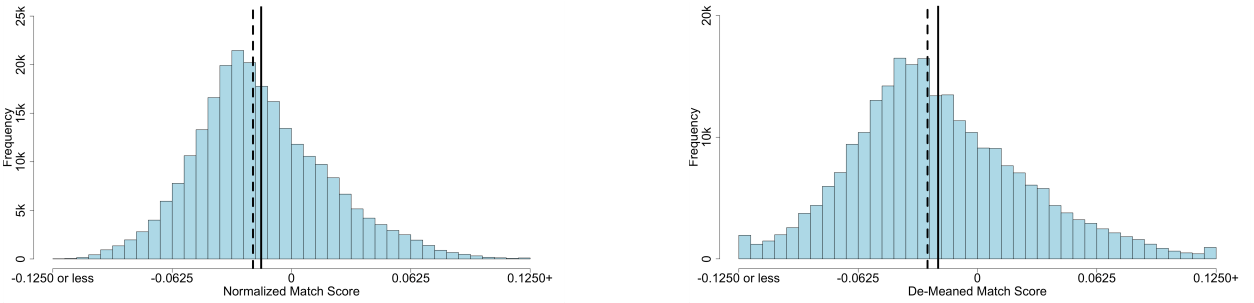
In the left image, the distribution of normalized author party match scores; in the right image, the distribution of de-meaned author-party match scores (without dividing by the entropy of the category distribution); the sample contains 264,672 opinions. Solid lines indicate the mean of each distribution (-0.014 and -0.010 respectively) and dashed lines indicate the medians (-0.014 and -0.009 respectively).

Figure C.25: Distributions of Panel Majority Match Scores



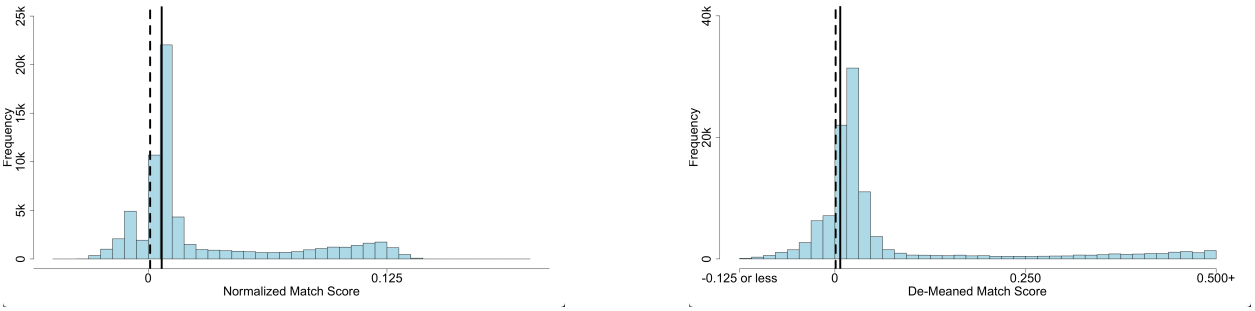
In the left image, the distribution of normalized panel majority match scores; in the right image, the distribution of de-meaned panel majority match scores (without dividing by the entropy of the category distribution); the sample contains 246,675 opinions. Solid lines indicate the mean of each distribution (-0.199 and -0.136 respectively) and dashed lines indicate the medians (-0.125 and -0.086 respectively).

Figure C.26: Distributions of Number-of-Democrats Match Scores



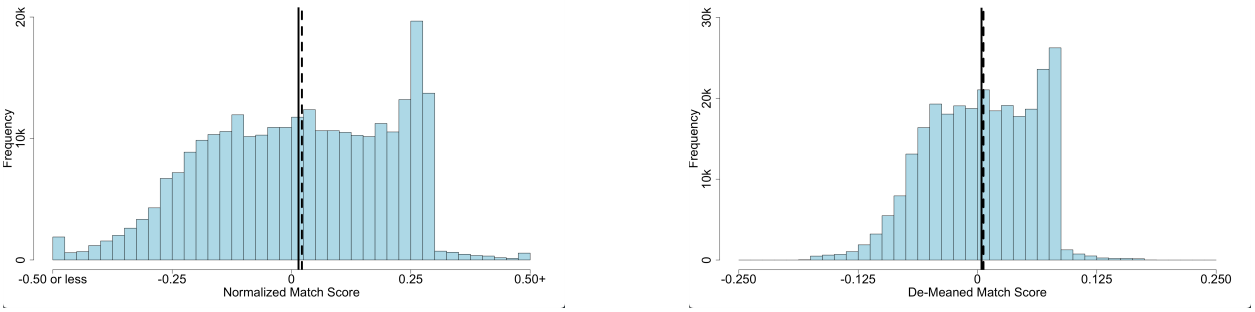
In the left image, the distribution of normalized number-of-Democrats match scores; in the right image, the distribution of de-meaned number-of-Democrats match scores (without dividing by the entropy of the category distribution); the sample contains 246,675 opinions. Solid lines indicate the mean of each distribution (-0.016 and -0.021 respectively) and dashed lines indicate the medians (-0.020 and -0.026 respectively).

Figure C.27: Distributions of Law School Match Scores



In the left image, the distribution of normalized law school match scores; in the right image, the distribution of de-meaned law school match scores (without dividing by the entropy of the category distribution). Solid lines indicate the mean of each distribution (0.007 and 0.027 respectively) and dashed lines indicate the medians (0.001 and 0.004 respectively). A majority of opinions have normalized match scores in $(-0.005, 0.005)$ and de-meaned match scores in $(-0.01, 0.01)$. For legibility, those opinions are omitted from the figure; the remaining samples contain 66,833 and 108,375 opinions respectively. 76,849 opinions have a normalized match score in $(-0.005, 0)$, and 130,033 opinions have a normalized match score in $(0, 0.005)$. 67,486 opinions have a de-meaned match score in $(-0.01, 0)$, and 97,854 opinions have a de-meaned match score in $(0, 0.01)$.

Figure C.28: Distributions of Gender Match Scores



In the left image, the distribution of normalized gender match scores; in the right image, the distribution of de-meaned gender match scores (without dividing by the entropy of the category distribution); the sample contains 274,158 opinions. Solid lines indicate the mean of each distribution (0.015 and 0.004 respectively) and dashed lines indicate the medians (0.022 and 0.006 respectively).

C.6.4 Robustness of Nearest-Neighbor Measures from Section 3.8.2

Table C.35: Opinion Similarity With 100 Neighbors

	(1) Topic	(2) Author Party	(3) Panel Majority	(4) # Dems.	(5) Law School	(6) Gender
Normalized match score	0.251	-0.009	-0.189	-0.012	0.009	0.016
	(0.00036)	(0.00028)	(0.00061)	(0.00021)	(0.00005)	(0.00164)
De-meaned match score	0.406	-0.006	-0.130	-0.015	0.033	0.005
	(0.00059)	(0.00019)	(0.00004)	(0.00026)	(0.00002)	(0.00049)
Share de-meaned > 0	0.879	0.464	0.241	0.397	0.596	0.572
Raw match share	0.638	0.499	0.378	0.277	0.080	0.852
Observations	270,451	264,672	246,675	246,675	273,715	274,158

All coefficients are significant at the 1% level.

Replicating Table C.11 using the 100 nearest neighbors (about 0.04% of the sample of opinions), instead of the 1,000 nearest neighbors as in that table. All analysis is conducted in the same way as in that table.

Table C.36: Opinion Similarity With 10,000 Neighbors

	(1) Topic	(2) Author Party	(3) Panel Majority	(4) # Dems.	(5) Law School	(6) Gender
Normalized match score	0.239	-0.020	-0.208	-0.020	0.005	0.018
	(0.00036)	(0.00019)	(0.00057)	(0.00018)	(0.00004)	(0.00150)
De-meaned match score	0.386	-0.014	-0.143	-0.025	0.018	0.005
	(0.00056)	(0.00013)	(0.00036)	(0.00022)	(0.00015)	(0.00045)
Share de-meaned > 0	0.883	0.358	0.133	0.201	0.687	0.546
Raw match share	0.618	0.491	0.365	0.267	0.066	0.853
Observations	270,451	264,672	246,675	246,675	273,715	274,158

All coefficients are significant at the 1% level.

Replicating Table C.11 using the 10,000 nearest neighbors (about 3.6% of the sample of opinions), instead of the 100 nearest neighbors as in that table. All analysis is conducted in the same way as in that table.

Table C.37: Opinion Similarity With 100 Neighbors, Excluding Own Author

	(1) Topic	(2) Author Party	(3) Panel Majority	(4) # Dems.	(5) Law School	(6) Gender
Normalized match score	0.251	-0.010	-0.191	-0.013	0.008	0.015
	(0.00036)	(0.00027)	(0.00061)	(0.00021)	(0.00005)	(0.00162)
De-meaned match score	0.405	-0.007	-0.131	-0.017	0.032	0.004
	(0.00006)	(0.00019)	(0.00039)	(0.00026)	(0.00002)	(0.00048)
Share de-meaned > 0	0.879	0.462	0.236	0.391	0.586	0.570
Raw match share	0.637	0.498	0.377	0.275	0.079	0.852
Observations	270,451	264,672	246,675	246,675	273,715	274,158

All coefficients are significant at the 1% level.

Replicating Table C.11, only considering neighbors with different authors. All analysis is conducted in the same way as in that table.

Table C.38: Opinion Similarity With 1,000 Neighbors, Excluding Own Author

	(1) Topic	(2) Author Party	(3) Panel Majority	(4) # Dems.	(5) Law School	(6) Gender
Normalized match score	0.248	-0.015	-0.201	-0.017	0.007	0.014
	(0.00035)	(0.00022)	(0.00057)	(0.00019)	(0.00004)	(0.00158)
De-meaned match score	0.400	-0.011	-0.138	-0.022	0.025	0.004
	(0.00058)	(0.00015)	(0.00037)	(0.00023)	(0.00018)	(0.00047)
Share de-meaned > 0	0.881	0.428	0.179	0.278	0.672	0.536
Raw match share	0.632	0.495	0.371	0.27	0.073	0.851
Observations	270,451	264,672	246,675	246,675	273,715	274,158

All coefficients are significant at the 1% level.

Replicating Table C.11 using the 1,000 nearest neighbors (about 0.4% of the sample of opinions), instead of the 100 nearest neighbors as in that table, and only considering neighbors with different authors. All analysis is conducted in the same way as in that table.

Table C.39: Opinion Similarity With 10,000 Neighbors, Excluding Own Author

	(1) Topic	(2) Author Party	(3) Panel Majority	(4) # Dems.	(5) Law School	(6) Gender
Normalized match score	0.251	-0.009	-0.189	-0.012	0.009	0.016
	(0.00035)	(0.00018)	(0.00056)	(0.00018)	(0.00003)	(0.00151)
De-meaned match score	0.406	-0.006	-0.13	-0.015	0.033	0.005
	(0.00054)	(0.00012)	(0.00036)	(0.00022)	(0.00014)	(0.00046)
Share de-meaned > 0	0.879	0.464	0.241	0.397	0.596	0.572
Raw match share	0.638	0.499	0.378	0.277	0.080	0.852
Observations	270,451	264,672	246,675	246,675	273,715	274,158

All coefficients are significant at the 1% level.

Replicating Table C.11 using the 10,000 nearest neighbors (about 3.6% of the sample of opinions), instead of the 100 nearest neighbors as in that table, and only considering neighbors with different authors. All analysis is conducted in the same way as in that table.

References

[1] Nadav Amir. Uniqueness of optimal strategies in captain lotto games. *International Journal of Game Theory*, vol. 47:no. 1, 2017. pp. 55-101.

[2] Anastasia Antsygina and Mariya Teteryatnikova. Optimal information disclosure in contests with stochasticprize valuations. *Economic Theory*, vol. 75:no. 1, 2023. pp. 743-780.

[3] Elliott Ash and Stephen Hansen. Text algorithms in economics. *Annual Review of Economics*, vol. 15:pp. 659–688, 2023. ISSN 1941-1391. doi:https://doi.org/10.1146/annurev-economics-082222-074352. URL https://www.annualreviews.org/content/journals/10.1146/annurev-economics-082222-074352.

[4] Elliott Ash, Daniel L. Chen, and Ariana Ornaghi. Gender attitudes in the judiciary: Evidence from u.s. circuit courts. *American Economic Journal: Applied Economics*, vol. 16 (1):pp. 314–350, 2024.

[5] Associated Press. Confidence in the supreme court remains low. *AP–NORC Center for Public Affairs Research*, 27 Jun. 2024:available online, 2024.

[6] Pak Hung Au and Keiichi Kawai. Competitive information disclosure by multiple senders. *Games and Economic Behavior*, vol. 119:pp. 56–78, 2020.

[7] Yakov Babichenko, Inbal Talgam-Cohen, and Konstantin Zabarnyi. Bayesian persuasion under ex ante and ex post constraints. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 5127–5134. Society for the Advancement of Artificial Intelligence, 2021.

[8] Michael R. Baye, Dan Kovenock, and Casper G. de Vries. The all-pay auction with complete information. *Economic Theory*, vol. 8:no. 2, 1996. pp. 291-305.

[9] Dorian Beauchêne, Jian Li, and Ming Li. Ambiguous persuasion. *Journal of Economic Theory*, vol. 179 (C):pp. 312–365, 2019.

[10] Robert M. Bell and Thomas M. Cover. Competitive optimality of logarithmic investment. *Mathematics of Operations Research*, vol. 5:no. 2, 1980. pp. 161-166.

[11] Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, vol. 57:no. 1, 2019. pp. 44-95.

[12] Peter Bils and Federica Izzo. Ideological spillovers and inefficient policy bundling. Working paper, 2022. URL https://drive.google.com/file/d/1PzsCzErXdV9URsfMTMEOiW_kuk-x0mWP/view.

[13] David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, vol. 24:no. 2, 1953. pp. 265-272.

[14] Raphael Boleslavsky and Christopher Cotton. Grading standards and education quality. *American Economics Journal: Microeconomics*, vol. 7:no. 2, 2015. pp. 248–279.

[15] Adam Bonica. Mapping the ideological marketplace. *American Journal of Political Science*, vol. 58 (2):pp. 367–386, 2014.

[16] Adam Bonica and Maya Sen. A common-space scaling of the american judiciary and legal profession. *Political Analysis*, vol. 25 (1):pp. 114–121, 2017.

[17] Isabelle Brocas and Juan D. Carillo. Influence through ignorance. *RAND Journal of Economics*, 38(4):931–947, 2007.

[18] Bernard Caillaud and Jean Tirole. Consensus building: How to persuade a group. *American Economic Review*, 97(5):1877–1900, 2007.

[19] Odilon Câmara and Jon X. Eguia. Slicing and bundling. *Journal of Politics*, 79(4): 1460–1464, 2017.

[20] Vinicius Carrasco, Vitor Farinha Luz, Nenad Kos, Matthias Messner, Paulo Monteiro, and Humberto Moreira. Optimal selling mechanisms under moment conditions. *Journal of Economic Theory*, vol. 177 (C):pp. 245–279, 2018.

[21] Vinicius Carrasco, Vitor Farinha Luz, Paulo Monteiro, and Humberto Moreira. Robust mechanisms: the curvature case. *Economic Theory*, vol. 68:no. 7, 2019. pp. 203–222.

[22] Cliff Carrubba, Barry Friedman, Andrew D. Martin, and Georg Vanberg. Who controls the content of supreme court opinions? *American Journal of Political Science*, vol. 56 (2):pp. 400–412, 2012.

[23] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages pp. 2898–2904. Virtual conference, nov 2020. doi:10.18653/v1/2020.findings-emnlp.261. URL https://aclanthology.org/2020.findings-emnlp.261.

[24] Yeon-Koo Che and Ian L. Gale. Caps on political lobbying. *The American Economic Review*, vol. 88:no. 3, 1998. pp. 643-651.

[25] Daniel L. Chen and Susan Yeh. Growth under the shadow of expropriation? the economic impacts of eminent domain. *Working paper*, cited version Mar. 2020:available online., 2020.

[26] Daniel L. Chen, Moti Michaeli, and Daniel Spiro. Non-confrontational extremists. *European Economic Review*, vol. 157:104521, 2023.

[27] Ying Chen and Hülya Eraslan. Informational loss in bundled bargaining. *Journal of Theoretical Politics*, 25(3):338–362, 2013.

[28] Ying Chen and Hülya Eraslan. Dynamic agenda setting. *American Economic Journal: Microeconomics*, 9(2):1–32, 2017.

[29] Alma Cohen. The pervasive influence of ideology at the federal circuit courts. *Working paper*, cited version Feb. 2024:available online, 2024.

[30] Alma Cohen and Rajeev H. Dehejia. Judges judging judges: Partisanship and politics in the federal circuit courts of appeals. *Working paper*, cited version Sep. 2024:available online., 2024.

[31] Joseph Copeland. Favorable views of supreme court remain near historic low. *Pew Research Center*, 8 Aug. 2024:available online, 2024.

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, Minneapolis, MN:pp. 4171–4186, 2019.

[33] Piotr Dworczak and Alessandro Pavan. Preparing for the worst but hoping for the best: Robust (bayesian) persuasion. *Econometrica*, vol. 90:no. 5, 2022. 2017-2051.

[34] Wioletta Dziuda. Strategic argumentation. *Journal of Economic Theory*, 146(4):1362–1397, 2011.

[35] Lee Epstein, Andrew D. Martin, Jeffrey A. Segal, and Chad Westerland. The judicial common space. *Journal of Law, Economics, and Organization*, vol. 23:pp. 303–325, 2007.

[36] Ky Fan. Minimax theorems. *Proceedings of the National Academy of Sciences*, vol. 39: no. 1, 1953. pp. 42-47.

[37] Chaim Fershtman. The importance of the agenda in bargaining. *Games and Economic Behavior*, 2(3):224–238, 1990.

[38] Alan Gelder, Dan Kovenock, and Brian Roberson. All-pay auctions with ties. *Economic Theory*, vol. 74:no. 4, 2022. pp. 1183-1231.

[39] Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. Measuring group differences in high-dimensional choices: Method and an application to congressional speech. *Econometrica*, vol. 87 (4):pp. 1307–1340, 2019.

[40] Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, vol. 18:no. 2, 1989. pp. 141-153.

[41] Micheal W. Giles, Virginia A. Hettinger, and Todd Peppers. Picking federal judges: A note on policy and partisan selection agendas. *Political Research Quarterly*, vol. 54 (3): pp. 623–641, 2001.

[42] Sergiu Hart. Discrete colonel blotto and general lotto games. *International Journal of Game Theory*, vol. 36:no. 3, 2008. pp. 441-460.

[43] Sergiu Hart. Allocation games with caps: from captain lotto to all-pay auctions. *International Journal of Game Theory*, vol. 45:no. 1, 2015. pp. 37-61.

[44] Ju Hu and Xi Weng. Robust persuasion of a privately informed receiver. *Economic Theory*, vol. 72:pp. 909–953, 2021.

[45] Ilwoo Hwang, Kyungmin Kim, and Raphael Boleslavsky. Competitive advertising and pricing. *Working paper*, 2022. last modified 30 August 2022, https://static1.squarespace.com/static/5271497de4b03475d0dd1240/t/636d516172b23e352e16883c/1668108642335/CAP_Draft_220830_Final.pdf.

[46] Matthew O. Jackson and Boaz Moselle. Coalition and party formation in a legislative voting game. *Journal of Economic Theory*, 103(1):49–87, 2002.

[47] Emir Kamenica. Bayesian persuasion and information design. *Annual Review of Economics*, vol. 11:no. 1, 2019. pp. 249-272.

[48] Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

[49] Jonathan P. Kastellec. Hierarchical and collegial politics on the u.s.courts of appeals. *The Journal of Politics*, vol. 73 (2):pp. 345–361, 2011.

[50] Jonathan P. Kastellec. Racial diversity and judicial influence on appellate courts. *American Journal of Political Science*, vol. 57 (1):pp. 167–183, 2013.

[51] Daniel M. Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J. Bommarito II. Natural language processing in the legal domain. *Working paper*, cited version Feb. 2023: available online., 2023.

[52] Andreas Kleiner, Benny Moldovanu, and Philipp Strack. Extreme points and majorization: Economic applications. *Econometrica*, vol. 89:no. 4, 2021. pp. 1557–1593.

[53] Anton Kolotilin, Tymofiy Mylovanov, Andriy Zapechelnyuk, and Ming Li. Persuasion of a privately informed receiver. *Econometrica*, vol. 85:no. 6, 2017. pp. 1949-1964.

[54] Svetlana Kosterina. Persuasion with unknown beliefs. *Theoretical Economics*, vol. 17:no. 3, 2022. pp. 1075-1107.

[55] Marie Laclau and Ludovic Renou. Public persuasion. *Working paper*, last modified 2 March 2016. https://core.ac.uk/download/pdf/48317828.pdf, 2016.

[56] Benjamin E. Lauderdale and Tom S. Clark. Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science*, vol. 58 (3):pp. 754–771, 2014.

[57] Maël Le Treust and Tristan Tomala. Persuasion with limited communication capacity. *Journal of Economic Theory*, 184:104940, 2019.

[58] Barton E. Lee. Gridlock, leverage, and policy bundling. *Journal of Public Economics*, 212:104687, 2022.

[59] Marin K. Levy and Adam S. Chilton. Challenging the randomness of panel assignment in the federal courts of appeals. *Cornell Law Review*, vol. 101 (1):pp. 1–56, 2015.

[60] Wei Lu and Daniel L. Chen. Motivated reasoning in the field: Polarization in precedent, prose, vote, and retirement in u.s. circuit courts, 1800-2013. *Working paper*, cited version Feb. 2024:available online, 2024.

[61] Fabio Maccheroni, Massimo Marinacci, and Aldo Rustichini. Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, vol. 74:no. 6, 2006. pp. 1447-1498.

[62] Helen Macdonald. *Vesper Flights*. Grove Atlantic, New York, New York, USA, 2020.

[63] Semyon Malamud and Andreas Schrimpf. Persuasion by dimension reduction. Working paper, 2022. URL https://arxiv.org/abs/2110.08884.

[64] Roger B. Myerson. Incentives to cultivate favored minorities under alternative electoral systems. *The American Political Science Review*, vol. 87:no. 4, 1993. pp. 856-869.

[65] Giri Parameswaran, Charles M. Cameron, and Lewis A. Kornhauser. Bargaining and strategic voting on appellate courts. *American Political Science Review*, vol. 115 (3):pp. 835–850, 2021.

[66] Keith T. Poole and Howard Rosenthal. A spatial model for legislative roll call analysis. *American Journal of Political Science*, vol. 29 (2):pp. 357–384, 1985.

[67] Pedro L. Rodriguez and Arthur Spirling. Word embeddings: What works, what doesn't, and how to tell the differencefor applied research. *Journal of Politics*, vol. 84 (1):pp. 101–115, 2024.

[68] Nicolas Sahuguet and Nicola Persico. Campaign spending regulation in a model of redistributive politics. *Economic Theory*, vol. 28:no. 1, 2006. pp. 95-124.

[69] Cass R. Sunstein, David Schkade, Lisa M. Ellman, and Andres Sawicki. Are judges political? an empirical analysis of the federal judiciary. *Brookings Institution Press*, 2006.

[70] Nora Szech. Tie-breaks and bid-caps in all-pay auctions. *Games and Economic Behavior*, vol. 92:pp. 138–149, 2015.

[71] Elias Tsakas and Nikolas Tsakas. Noisy persuasion. *Games and Economic Behavior*, 130: 44–61, 2021.

[72] Alexander Wolitzky. Mechanism design with maxmin agents: Theory and an application to bilateral trade. *Theoretical Economics*, vol. 11:no. 3, 2016. pp. 971–1004.