Institutional Design of Criminal Justice Processes

Thesis by
Joanna Nanami Huey

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2024
Defended May 13, 2024

Joanna Nanami Huey
ORCID: 0009-0000-5934-9184

For David

ACKNOWLEDGEMENTS

This work never would have begun, and certainly never would have finished, without my advisor Alex Hirsch. His class in my first year at Caltech opened up ways of thinking about the world, setting me on a new path. His steadfast efforts as a mentor have guided me along that path and pushed me to keep going when I stumbled. I am grateful for the countless hours he has spent engaged with these projects and for his support in balancing research with the other demands of life. I hope eventually to emulate his rigorous standards for reasoning, his clear presentation of complicated ideas, and his dedication to his students. I also hope that his future students have more interest in coffee, homemade yogurt, and reasonable speeds during walking meetings, and less interest in the generation of administrative hurdles.

I am deeply thankful for my committee members: Jonathan Katz, Tom Palfrey, and Jean-Laurent Rosenthal. They bring different perspectives, but all have an uncanny ability to see to the heart of a problem, asking the important questions and revealing the potential in a line of research. When I am stuck or lost in the weeds, I am fortunate to be able to draw upon their insights and experience.

Discussions with the applied political theory group sharpened my thinking. Many thanks to Saba Devdariani, Lindsey Gailmard, Mike Gibilisco, and Jacob Morrier for sitting through muddled presentations and weekly complaints. You all have been my favorite research commitment device.

The Caltech HSS faculty have been extraordinarily generous with time and advice. I am particularly grateful to Laura Doval, Gabriel López-Moctezuma, Kirby Nielsen, Bob Sherman, and Omer Tamuz. Having been in school for longer than anyone ought to be, I am lucky to have learned from exemplary teachers and researchers. Many thanks to Nick Feamster, Ed Felten, Howard Georgi, Bruce Mann, William Rubenstein, and Steven Shavell. Thanks also to Susannah Barton Tobin, who gave extensive advice about the job market to a not-at-all-recent alum.

Laurel Auchampaugh and Kapaūhi Stibbard keep so many things running in the department. I have at least 808 email threads from Laurel, ranging from the one admitting me to the program to the one reminding me to submit this thesis, all steering me with endless patience and understanding. Despite my antagonistic relationship with the Baxter A/V systems, Kapaūhi was always able to sort out my last-minute problems with good humor.

Work is better with friends. My starting cohort — Sumit Goel, Daniel Guth, Wade Hann-Caruthers, and Jeff Zeidel — got me through years of classes, research, teaching, and general uncertainty. My job market cohort — Kate Huang, Shunto Kobayashi, Po-Hsuan Lin, and Aldo Lucia — got me through one very particular uncertainty. Thanks to everyone in the graduate proseminar for listening, with particular thanks to Peter Doe, Danny Ebanks, and Matt Estes for their comments. I am looking forward to seeing the amazing things you all do. Thanks also to Jason Abaluck for convincing me to apply in the first place; I did not "enroll and then drop out because they made [me] take a stupid theory class," so I think I held up my end of the bargain.

I am immensely glad to have had my family in Southern California during my time at Caltech. My mom and dad, Nancy Nakasone-Huey and Don Huey, and my in-laws, Bev Simmons and Ross Duffin, have kept us sane, well-fed, and loved. Thank you so much for your decades of support and for always believing the best of me. My children, Miko and Naomi, constantly remind me of the joy of learning, the importance of snacks, and the fact that my research is not terribly interesting to the 7-and-under set. I love you all.

Finally, as is only just, I am here to give a thesis for a thesis. This work is dedicated to my husband, David Simmons-Duffin, who has juggled so much over these past years. It is not easy to be a new assistant professor with a toddler and a partner who stays up until 3 am trying to remember how to do problem sets. It is not easy to live in the constant sleep deprivation of a night owl responsible for the morning school run. It is not easy to be a good father, a good husband, or a good physicist, and yet you manage to do all three. Thank you for doing the difficult things. Like Matt Gline, I sure picked a good lab partner.

ABSTRACT

This dissertation contains three essays that contribute to ongoing debates about the design of institutions and procedures related to criminal justice.

Chapter 1 investigates how peremptory challenges in the jury selection process affect the diversity of and outcomes from juries. A game-theoretic model of attorneys' decisions to strike potential jurors finds that the process 1) can lead selected jurors from a majority group to be a skewed sample and 2) can increase minority representation, contrary to common intuition. The first theoretical finding about the skew is supported by empirical analysis of data from jury selection transcripts: a novel measure of the pro-defense lean of jury pool members is developed, and selected White jurors are found to be more pro-defense than the average White pool member.

Chapter 2 develops a game-theoretic model of decisions about the verdict and sentence in a criminal trial, considering both single-actor and two-actor versions of this two-step process. Restrictions on sentencing discretion can lead to nullification where an actor with acquits who would have convicted under full discretion. When actors care about the lawfulness of their own actions, a two-actor process may lead to additional convictions, as the convicting actor can free ride off of a separate sentencing actor who will pay the cost of sentencing away from the lawful sentence. The model also leads to non-monotonic effects on the verdict when lawfulness or the expected sentence change.

Chapter 3 (joint work with Alexander V. Hirsch) uses mechanism design to examine single-threshold information escrows in a workplace setting. In this setting, reports of misconduct by a manager are kept secret until the number of reports exceeds a threshold and the manager is fired. When the firm designing the system wishes to minimize misconduct, a single-threshold mechanism leads to optimal results when misconduct reports are costless. In contrast, costly misconduct reports can make truthful reporting impossible under certain threshold values, raising the threshold above the firm's ideal or even eliminating the possibility of any truthful mechanism. We find that single-threshold mechanisms are generally worse for the firm than mechanisms that mix two thresholds and can be worse than choosing whether to fire the manager without eliciting any information about misconduct.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

While certain criminal justice institutions and procedures have garnered extensive theoretical analysis, many aspects remain less examined. This dissertation contains three essays considering such aspects and contributing to ongoing debates about the design of these institutions and procedures. Chapters 1 and 2 study criminal trial processes. Chapter 1 analyzes the effects of design choices in the jury selection process on the composition of juries, using a game theoretic model of attorneys' decisions to strike potential jurors and empirical analysis of a dataset including strike decisions and jury pool member characteristics. Chapter 2 considers the effects of having different people determine the conviction and sentence, modeling the consequences of strategic choices by a jury deciding on a defendant's guilt knowing that a judge will sentence the defendant if found guilty. Chapter 3 studies a potential supplement or alternative to criminal trials: information escrows in which reports of misconduct are kept secret until a threshold number of reports is exceeded. In this chapter, a mechanism design approach allows consideration of when such escrows can yield truthful reporting, and how the threshold choice affects the designer's expected utility.

Chapter 1 investigates how selection procedures affect the diversity of and outcomes from juries. Attorneys' culling of jurors through peremptory challenges is believed to create unrepresentative and harsh juries. However, this chapter constructs a model showing that peremptories can lead to (i) a jury that is more pro-defense than it looks and (ii) a jury that overrepresents a minority group. The model includes a heterogenous majority group (split into pro-defense and pro-prosecution subgroups) and a homogenous minority group. This key feature yields two main results. First, when the minority is far from the majority's mean, more majority jurors close to the minority are selected, and this skew can increase acquittals. Empirically, I use jury selection transcripts to construct a novel measure of pro-defense lean and confirm the skew: selected White jurors are more pro-defense than the average White pool member. Second, when the minority lies between the majority's subgroups, peremptories actually increase minority representation. This overrepresentation helps to explain earlier conflicting empirical findings on how peremptories affect jury demographics. Together, these results indicate that recent and proposed reforms to eliminate or alter peremptory challenges may not consistently improve representation or reduce convictions.

Chapter 2 also uses game theory to model the interactions between strategic actors involved in criminal trials. I model decisionmaking about the verdict and sentence, which can be decided by a sequence of two actors — typically a jury and a judge — or by a single actor. First, results formalize the idea that restrictions on sentencing discretion, either from a mandatory minimum or from another actor controlling the sentence, leads to a type of jury nullification where the jury acquits even though it would have convicted with full sentencing discretion. Second, I extend the model to consider actors who care not only about whether mistaken outcomes occur (convicting the innocent, acquitting the guilty) but also about whether they are obeying the law. With the addition of this lawfulness utility, the actor deciding on a verdict may free ride off of the actor deciding the sentence if the two actors have similar ideal sentences. In other words, an actor can acquit when responsible for both steps of the process, but convict when another actor needs to pay the cost of sentencing away from the lawful sentence. In addition, I find non-monotonic effects on the verdict choice when an actor's lawfulness is increased or when the expected sentence is increased.

Chapter 3 — joint work with Alexander V. Hirsch — takes a mechanism design approach to considering how to engineer a type of reporting system called an information escrow to improve reporting of misconduct. Information escrows currently are used to reduce the first-mover disadvantage for reporting sexual harassment by keeping reports secret until more than one report names the same perpetrator. We consider a firm seeking to minimize misconduct using a single-threshold mechanism, where the firm commits to fire a manager when employees report more misconduct incidents than the threshold value. When reporting is costless, the firm can attain its first-best result, eliciting truthful reports using the firm's own ideal threshold. However, when misconduct reports are costly, truthful reporting occurs only when the reporting cost falls between an employee's net benefit of falsely reporting misconduct and the net benefit of truthfully reporting misconduct. These net benefits depend on the employee being pivotal to the decision, and pivotality depends on the threshold. So, the range of reporting costs supporting truthful reporting varies with the threshold, and in fact there may be certain thresholds for which no costs permit truthful reporting and certain costs for which no thresholds support truthful reporting. From the firm's perspective, even if truthful, single-threshold mechanisms exist for a given reporting cost, they may be dominated by making the firing decision solely based on the firm's prior belief about the manager or by mechanisms that mix two different thresholds.

All three chapters are motivated by the idea that modeling strategic choices by actors in criminal justice processes can yield insights that neither intuition nor empirical study alone have made obvious. In Chapter 1, the theoretical prediction of the potential for overrepresentation of a minority group on juries squares with empirical studies having mixed results on the effect of peremptories on representation. The studies finding increased representation, as opposed to the expected decrease, generally characterized those results as surprising and without explanation. The model here gives a testable theory about when we should expect to see over- and underrepresentation. This chapter also offers an example of how theory can influence empirical work, as the focus of the empirical analysis was suggested by the theoretical prediction of a skew in selected majority jurors. Chapter 2 produces two counterintuitive results: 1) the potential for a two-actor process to increase convictions because of lenient juries free-riding off of lenient judges willing to lower sentences, and 2) the potential for increasing lawfulness to move the trial outcome from conviction (with a low, unlawful sentence) to acquittal to conviction (with a high, lawful sentence) and the potential for increasing sentences to move the trial outcome from acquittal, to conviction (as the sentence approaches the ideal of the actor determining guilt), to acquittal again. Chapter 3 demonstrates that the single-threshold mechanisms currently used for information escrows may not be compatible with truthful reporting for certain reporting costs, and may yield worse outcomes than mechanisms with mixed thresholds or even than simply using the prior beliefs about the manager and eliciting no reports at all. I hope that these projects offer some steps forward in the literature connecting theoretical and empirical analysis of legal institutions.

Chapter 1

THE EFFECTS OF PEREMPTORY CHALLENGES ON JURY DIVERSITY AND CONVICTION RATES

Peremptory challenges allow attorneys to strike potential jurors without stating any cause, opening a back door for discrimination. Many worry that peremptories lead to the underrepresentation of marginalized groups on juries and that this underrepresentation increases conviction rates. For example, Liptak (2015) discusses prosecutors challenging disproportionate numbers of Black jury pool members, citing a study finding lower conviction rates with more Black jurors and quoting an expert: "'If you repeatedly see all-white juries convict African-Americans, what does that do to public confidence in the criminal justice system?"' In Scapicchio (2022), a defense attorney argues for eliminating prosecutorial peremptory challenges, stating that "[n]on-diverse juries are more likely to convict Black and brown defendants."

These concerns have led to decades of calls to eliminate or reform peremptories. In Batson v. Kentucky (1986) — a Supreme Court case holding that peremptory challenges based solely on race violate the Equal Protection Clause — Justice Thurgood Marshall's concurrence argued that the "decision today will not end the racial discrimination that peremptories inject into the jury selection process. That goal can be accomplished only by eliminating peremptory challenges entirely." (102-03) States have started to answer these calls.[1] In 2018, the Washington Supreme Court adopted a new rule (Wash. Gen. R. 37) to make it easier to challenge peremptories for being discriminatory, and California enacted a similar law in 2020 (A.B. 3070). Going further, the Arizona Supreme Court adopted changes to court rules that eliminated peremptory challenges beginning in 2022. (Ariz. R. Crim. P. 18; Ariz. R. Civ. P. 47)

This paper constructs a model, supported by empirical analysis, that questions whether eliminating peremptories will consistently improve representation and reduce convictions. The model considers two strategic attorneys who each have one

[1]The Death Penalty Clinic at UC Berkeley School of Law has compiled an extensive list of reforms and proposed reforms to peremptory challenges. https://www.law.berkeley.edu/experiential/clinics/death-penalty-clinic/projects-and-cases/whitewashing-the-jury-box-how-california-perpetuates-the-discriminatory-exclusion-of-black-and-latinx-jurors/batson-reform-state-by-state/

strike to use in choosing a single-juror jury from a pool of three potential jurors in a known order. As such, the attorneys decide whether to strike the potential juror currently being questioned based in part on their knowledge of the next ones in line. While the details of jury selection procedures vary widely between jurisdictions, this type of strike decision does occur regularly. Munsterman et al. (2006) warns: "If by waiving its remaining peremptory challenges, a party effectively excludes a previously identified juror (e.g., the parties have access to the randomized list and can identify the next juror to be called into the jury box for questioning), the waiver can be challenged under *Batson*." A trial and jury consulting firm advises: "Asking the court for the randomized order of the juror list is critical because this randomized order of the jurors will affect the strategy you take in selecting (and striking) jurors. For instance, your worst jurors may be at the end of the panel, in which case you may choose to pass on your strikes early in the selection process" (Trial Partners, 2012).

The model incorporates intrinsic preferences of jurors that make them more or less pro-prosecution. It includes a heterogenous majority group, split into relatively pro-defense and pro-prosecution subgroups. It also includes a homogenous minority group, unified in its amount of pro-prosecution lean. This lean can be at any level relative to the majority group's, including at an intermediate level between the two majority subgroups. Such an "intermediate" minority might be expected if attitudes about prosecution correlate with preferences for police funding. In Pew Research Center (2021), 49% of White Americans stated that spending on policing in their area should be increased, compared to 38% of Black Americans. However, only 32% of White Democrats desired increased spending, compared to 38% of Black Democrats and 64% of White Republicans. In general, allowing this heterogeneity within the majority group as well as heterogeneity between the majority and minority groups yields two results.

First, a minority group that is not intermediate will be underrepresented, but the majority group members who are more similar to the minority will be more likely to be selected as jurors. More concretely, consider a jury pool with two majority members and one minority member who favors acquittal more than both majority members. If the prosecutor used a strike on the minority member, the defense attorney may be free to strike a conviction-leaning majority member, resulting in the more acquittal-leaning majority member becoming the jury. Similar to an effect hypothesized in Anwar et al. (2012), this skew in selected majority jurors stems

from a gravitational force of "invisible" minority members, who are in the pool but not the jury. Such a skew can be large enough to outweigh the effect of minority underrepresentation on conviction rates, in which case acquittal is more likely than in a system without peremptories.

Second, an intermediate minority group actually will be overrepresented. Because peremptories tend to remove extreme jurors, if the majority jury pool members are polarized and the minority jury pool members lie between those poles, then minority jurors are selected disproportionately. Again, acquittal is more likely than without peremptories.

These two results arise both in a baseline version of the model and in a variant, which reflect the most commonly used peremptory challenge procedures. The baseline model follows a "strike-and-replace" procedure. Initially, attorneys know whether each potential juror belongs to the minority or majority group. When a potential juror is questioned, attorneys learn her precise pro-prosecution lean and decide whether to strike her before questioning the next potential juror. In contrast, the model variant reflects the "struck jury" method. Before making any strike decisions, the attorneys already know the pro-prosecution lean for all three potential jurors. Both model versions predict a skew in majority jurors for a minority that is not intermediate. However, the strike-and-replace version also predicts such a skew for an intermediate minority that is similar enough to one of the majority's subgroups. In that case, an attorney may decide to strike a minority jury pool member, gambling that unquestioned potential majority jurors later in the order will turn out to favor their side.

Empirically, this paper uses an unusually detailed dataset from Craft (2018) to construct a proxy measure of pro-prosecution lean and confirm the skew described in the first theoretical result. Investigative reporters gathered data from the Fifth Circuit Court District of Mississippi. The data includes demographic information for jury pool members, as well as hand-coded binary descriptors of pool members responses to questions posed to the entire pool. This rare level of information gives insight into the intrinsic preferences of each potential juror, using the same responses that the attorneys used when making their strike decisions.

The proxy measure is count-based, subtracting the number of pro-defense responses from the number of pro-prosecution responses for each pool member. The measure aligns with the actual peremptory challenge decisions made by the attorneys, meaning that pool members who are more pro-prosecution according to the proxy

are more often struck by defense attorneys and those who are less are more often struck by prosecutors. Analyzing the dataset, White jury pool members are more pro-prosecution than Black jury pool members. Furthermore, White selected jurors are significantly less pro-prosecution than White jury pool members. This work provides the first direct evidence of a skew in selected majority jurors.

1.1 Related Literature

One distinguishing feature of this model is its treatment of the intrinsic preferences of jurors, which determine how likely they are to convict. Having heterogeneous preferences within a group allows for assessment of what kinds of jurors get chosen from a particular group. Permitting the minority and majority group to have different preferences allows for analysis of different kinds of jury pools: the relative pro-prosecution leans of minority and majority pool members may vary based on the defendant's identity, the type of crime, or the local demographics.

Most earlier models of peremptories take all jurors to come from a single population-wide distribution of preferences (Roth et al., 1977; Brams & Davis, 1978; DeGroot & Kadane, 1980; Kadane et al., 1999; Flanagan, 2015). Four exceptions are Neilson & Winter (2000), Lehmann & Smith (2013), Moro & Van der Linden (2022), and Anwar et al. (2012). Neilson & Winter (2000) model two groups, but assume that each group is homogenous, with all group members having the identical likelihood of conviction. Lehmann & Smith (2013) model a population with one pro-prosecution type and one pro-defense type, split into two groups that have different proportions of those types. Moro & Van der Linden (2022) allow for generalized distributions of types, but run simulations under the assumption that one distribution first-order stochastically dominates the other. The distribution assumptions in these three papers foreclose some or all of the effects found in this paper.

Though primarily empirically focused, Anwar et al. (2012) contains a theoretical example with normal distributions of conviction likelihoods for Black and White jurors. They assume that the tails of these distributions will be removed by peremptory challenges, which parallels the attorney behavior in the struck-jury version of the current model. So, they find a skew in selected jurors similar to the one theorized here, and the proxy measure described above demonstrates that this theorized skew actually occurs. This paper also extends the theoretical analysis by evaluating a strike-and-replace setting, as well as by considering the possibility of an intermediate minority and the resulting possibility of minority overrepresentation.

Although no prior theoretical work has predicted a region of minority underrepresentation and skew and a complementary region of overrepresentation, both underrepresentation and skew have appeared as outcomes. Two simulation studies predict underrepresentation. Ford (2009) argues that strikes induce a tendency towards the median that overrepresents the majority. Revesz (2016) simulates strikes based entirely on demographic stereotyping and finds a decrease in the number of Democratic jurors chosen because they are easier to identify precisely and strike. In addition to the skew in Anwar et al. (2012) discussed above, Schwartz & Schwartz (1996) also describes a hypothetical example in which the proportion of men and women on a jury remains the same as in the pool, but the extremes of each gender group are struck.

Unlike underrepresentation, overrepresentation of a minority group through peremptory challenges does not appear in the theoretical literature.[2] However, empirical work suggests it occurs. Anwar et al. (2012, 1030) see an increase from 3.9% (Black members of the jury pool) to 4.6% (Black members of the jury). Rose (1999) finds a higher percentage of White potential jurors (49%) struck through peremptories than Black potential jurors (42%). Similarly, Gau (2016) finds 35% of White potential jurors struck through peremptories, as compared to 30% of Blacks, 26% of Hispanic/Latinos, 29% of Asians, and 21% of other races.

The theoretical literature also does not suggest that peremptories may increase acquittals. The mixed effect on conviction rates predicted by this model may help to explain why many defense attorneys lobby for the retention of peremptories. Public defenders have opposed legislation reducing the number of peremptories, and defense attorneys call peremptory strikes "invaluable" (Leshem, 2019, n.56).

The simplification of selecting a single-member jury means that this model does not speak to the questions of jury size (Brams & Davis, 1978; Flanagan, 2015), likelihood of unanimity (Schwartz & Schwartz, 1996), or strategic juror interaction (e.g., Austen-Smith & Banks, 1996; Feddersen & Pesendorfer, 1998; Duggan & Martinelli, 2001). However, the model's results on the distribution of the selected juror's group membership and the expected conviction rate of the jury relate to prior theoretical and empirical results on the composition and conviction rates of multi-member juries. In addition, this paper relates, more loosely, to work on selection by committee members with vetoes (Alpern & Gal, 2009; Alpern et al., 2010), the

[2]Lehmann & Smith (2013) does allow for minority overrepresentation if a defense attorney is highly skilled at convincing judges to strike jurors *for cause*, a type of strike separate from peremptory challenges.

selection of an arbitrator using veto-rank and shortlisting (De Clippel et al., 2014), and a mechanism design approach to struck-jury procedures (Van der Linden, 2017).

1.2 Model

Table 1.1: Notation

D	defense attorney
P	prosecutor
i	potential jurors $\in \{1, 2, 3\}$
g_i	group $\in \{A, B\}$
ω	state $\in \{0, 1\}$, indicating if defendant is not guilty or guilty
v_i	juror i's vote $\in \{0, 1\}$, indicating acquittal or conviction if juror i is selected
s_i	juror i's signal, $s_i \in (0, \infty)$
π_i	juror i's type
α	proportion of population belonging to group A
ρ	prior probability of guilt $\in (0, 1)$
ν, γ	$f(s_i\|\omega = 0) = \nu e^{-\nu s_i}$ $f(s_i\|\omega = 1) = \gamma e^{-\gamma s_i}$ $\nu > \gamma > 0, \frac{\gamma}{\nu} < \frac{\pi_i}{1-\pi_i}\frac{1-\rho}{\rho}$
x_ω	$= \nu$ if $\omega = 0$ or γ if $\omega = 1$
$\underline{a}, \overline{a}, \underline{b}, \overline{b}$	possible values of π_i; $\mathbb{P}[\pi_i = \underline{a}\|g_i = A] = \mathbb{P}[\pi_i = \overline{a}\|g_i = A] = \mathbb{P}[\pi_i = \underline{b}\|g_i = B] = \mathbb{P}[\pi_i = \overline{b}\|g_i = B] = 0.5$
$c(\omega, \pi_i)$	probability of conviction given a juror of type π_i
$c(\omega, g_i)$	probability of conviction given a juror from group g_i

Two attorneys, D and P, select a single-member jury for the trial of a defendant who the attorneys know is guilty ($\omega = 1$) or not ($\omega = 0$). A selected juror i will vote to acquit ($v_i = 0$) or convict ($v_i = 1$).

Underlying Juror Behavior

The jury pool has three potential jurors, $i \in \{1, 2, 3\}$, each of whom belongs to one of two groups, $g_i \in \{A, B\}$. Unlike the attorneys, the jurors do not know the guilt of the defendant, but know that defendants have a prior probability of guilt $\rho \in (0, 1)$. If selected to be the jury, each juror i has state-dependent payoffs, with different losses for false acquittals and false convictions, both determined by the juror's type

π_i:

	$\omega = 0$	$\omega = 1$
$v_i = 0$	0	$-(1-\pi_i)$
$v_i = 1$	$-\pi_i$	0

The groups A and B each contain two types of jurors, and so each group has a different, two-point distribution of π_i:

$$\mathbb{P}[\pi_i = \underline{a}|g_i = A] = \mathbb{P}[\pi_i = \overline{a}|g_i = A] = 0.5$$
$$\mathbb{P}[\pi_i = \underline{b}|g_i = B] = \mathbb{P}[\pi_i = \overline{b}|g_i = B] = 0.5$$

where $\underline{a}, \overline{a}, \underline{b}, \overline{b} \in (0,1)$, $\underline{a} \leq \overline{a}$, and $\underline{b} \leq \overline{b}$.

The selected juror bases her vote v_i on these π_i-dependent payoffs as well as on a signal $s_i \in (0, \infty)$ obtained during the trial. The signal has different exponential distributions depending on the state, with probability distribution functions as follows:[3]

$$f(s_i|\omega = 0) = \nu e^{-\nu s_i}$$
$$f(s_i|\omega = 1) = \gamma e^{-\gamma s_i}$$

The juror will vote to convict when the conditional probability of guilt given the signal is greater than π_i:

$$\begin{aligned} EU_i(v_i = 1) &\geq EU_i(v_i = 0) \\ \mathbb{P}[\omega = 0|s_i](-\pi_i) &\geq \mathbb{P}[\omega = 1|s_i](-(1-\pi_i)) \\ \mathbb{P}[\omega = 1|s_i] &\geq \pi_i \end{aligned}$$

Note that this condition is equivalent to

$$\begin{aligned} \frac{\mathbb{P}[s_i|\omega = 0]\mathbb{P}[\omega = 0]}{\mathbb{P}[s_i]}(-\pi_i) &\geq \frac{\mathbb{P}[s_i|\omega = 1]\mathbb{P}[\omega = 1]}{\mathbb{P}[s_i]}(-(1-\pi_i)) \\ \frac{\gamma e^{-\gamma s_i}}{\nu e^{-\nu s_i}} &\geq \frac{\pi_i}{1-\pi_i}\frac{1-\rho}{\rho} \end{aligned} \tag{1.1}$$

Assumption 1. *(Signal Monotonicity)*

$$\nu > \gamma > 0$$

[3]These distributions appear in an example in Duggan & Martinelli (2001).

This assumption means that a higher signal means a higher likelihood of guilt. It also implies that $\frac{\gamma e^{-\gamma s_i}}{\nu e^{-\nu s_i}}$ on the left-hand side of equation 1.1 increases with s_i, and the juror will vote to convict whenever the trial signal is greater than or equal to the s_i^* that satisfies equation 1.1 with equality. In other words, the juror votes to convict if and only if $s_i \geq s_i^*$.

Assumption 2. *(Possibility of Acquittal)*

$$\frac{\gamma}{\nu} < \frac{\pi_i}{1-\pi_i}\frac{1-\rho}{\rho}$$

This assumption implies that there exists some small enough $s_i \in (0, \infty)$ such that the juror will acquit. Note that there will always be a value of s_i such that the juror convicts since $\frac{\gamma e^{-\gamma s_i}}{\nu e^{-\nu s_i}} \to \infty$ as $s_i \to \infty$.

Definition 1. *Label the probability that juror i convicts as a function of the state ω and the juror's conviction threshold π_i as*

$$\boldsymbol{c(\omega, \pi_i)} \equiv \mathbb{P}[s_i > s_i^* | \omega, \pi_i]$$

and the probability that a juror from group A convicts as

$$\boldsymbol{c(\omega, A)} \equiv \mathbb{P}[s_i > s_i^* | \omega, g_i = A] = \frac{1}{2}c(\omega, \underline{a}) + \frac{1}{2}c(\omega, \overline{a})$$

with the parallel definition for $c(\omega, B)$.

Attorney Preferences

When selecting the single-member jury, the attorneys care solely about the selected juror's eventual vote to acquit or convict, $v \in \{0, 1\}$:

$$u_D = \begin{cases} 1 & \text{if } v = 0 \\ 0 & \text{if } v = 1 \end{cases}$$

$$u_P = \begin{cases} 0 & \text{if } v = 0 \\ 1 & \text{if } v = 1 \end{cases}$$

Since the attorneys' payoffs are each 0 for one value of v and 1 for the other value of v, the expected payoff for the prosecutor is simply the probability that the selected juror convicts conditional on the state. For notational simplicity, define

$$x_\omega = \begin{cases} \nu & \text{if } \omega = 0 \\ \gamma & \text{if } \omega = 1. \end{cases}$$

Then, the expected payoff for the prosecutor is

$$c(\omega, \pi_i) = \int_{s_i^*}^{\infty} x_\omega e^{-x_\omega s_i}\, ds_i \; = e^{-x_\omega s_i^*}$$

and the payoff is

$$1 - c(\omega, \pi_i) = 1 - e^{-x_\omega s_i^*}$$

for the defense attorney.

Sequence: Strike-and-Replace (Attorneys Know Groups, Learn Types)
Jury selection proceeds via a simplified version of the strike-and-replace method used in the majority of U.S. courts. The attorneys each have one strike to veto a potential juror, and initially they know the order of the three potential jurors as well as the group membership of each of the three—imagine a jury pool seated in order in the courtroom, where attorneys can guess about upcoming jurors based on what they see. The sequence proceeds as follows:

1. Nature chooses the guilt of the defendant ω and the three jurors' payoffs π_1, π_2, π_3, which fully determine the jurors' conviction thresholds s_1^*, s_2^*, s_3^*.
2. The attorneys learn ω and the jurors' group memberships g_1, g_2, g_3; jurors know only the prior probability of guilt ρ.
3. Until a juror is selected to be the jury, for each juror in order, any attorneys with a remaining strike question the juror, learn π_i, and then simultaneously choose whether to strike the juror, and
    - if neither attorney strikes the juror, the juror becomes the jury,
    - any attorney who uses a strike can never strike again, and
    - if an attorney strikes the juror, the process repeats for the next juror in order.
4. A trial occurs, and the selected juror obtains a signal s and votes to acquit or convict ($v = 0$ or $v = 1$), and payoffs are realized for the selected juror and the attorneys.

1.3 Attorney Behavior in Equilibrium

Deciding to Strike Juror 2
Working backwards through the strike decisions to find the subgame perfect equilibrium, first consider the last decision made by the attorneys: whether to strike juror

2. Note that an attorney only reaches this decision if juror 1 has been struck by his opponent, using up the opponent's strike, so at most one attorney will be making this strike decision.

The attorney will examine juror 2, learn π_2 and calculate the value of juror 2's threshold

$$\frac{\gamma e^{-\gamma s_2^*}}{\nu e^{-\nu s_2^*}} = \frac{\pi_2}{1-\pi_2}\frac{1-\rho}{\rho}$$
$$s_2^* = \frac{1}{\nu-\gamma}\ln\left(\frac{\nu}{\gamma}\frac{\pi_2}{1-\pi_2}\frac{1-\rho}{\rho}\right)$$

Then, the expected payoff for the prosecutor of juror 2 being selected is

$$c(\omega, \pi_2) = \left(\frac{\nu}{\gamma}\frac{\pi_2}{1-\pi_2}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}}$$

and the corresponding expected payoff for the defense attorney is $1 - c(\omega, \pi_2)$.

The attorney would compare this expected payoff for choosing juror 2 to the expected payoff for striking juror 2 and thereby choosing juror 3, for whom only the group g_3 and not π_3 is known. For the prosecutor, these expected payoffs are

$$c(\omega, g_3) = \begin{cases} \frac{1}{2}\left(\frac{\nu}{\gamma}\frac{\underline{a}}{1-\underline{a}}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}} + \frac{1}{2}\left(\frac{\nu}{\gamma}\frac{\overline{a}}{1-\overline{a}}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}} & \text{if } g_3 = A \\ \frac{1}{2}\left(\frac{\nu}{\gamma}\frac{\underline{b}}{1-\underline{b}}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}} + \frac{1}{2}\left(\frac{\nu}{\gamma}\frac{\overline{b}}{1-\overline{b}}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}} & \text{if } g_3 = B \end{cases}$$

The defense attorney's expected payoffs are 1 minus the prosecutor's expected payoffs in each case.

Deciding to Strike Juror 1

Next, consider the decision made by both attorneys about whether to strike juror 1. Here, each attorney knows that, if he uses his strike, the other attorney will have the opportunity to use her strike after examining juror 2 and learning s_2^*. If neither attorney strikes, juror 1 will be the jury. If both attorneys strike, juror 2 will be the jury.

In general, the attorney payoffs are as in table 1.2, where v_i is the expected value of the vote of juror i given the distribution of s_i.

If one attorney strikes, it is always (weakly) better for the other attorney not to strike because by saving his strike, the attorney can then choose the better of juror 2 (after π_2 is known) and the expectation of juror 3 based on g_3, rather than being forced to keep

Table 1.2: Attorney Payoffs

		Prosecutor: Strike	Prosecutor: No Strike
Defense	Strike	$\mathbb{E}[u_D(v_2)]$, $\mathbb{E}[u_P(v_2)]$	$\mathbb{E}[u_D(\max\{v_2, \mathbb{E}[v_3]\})]$, $\mathbb{E}[u_P(\max\{v_2, \mathbb{E}[v_3]\})]$
	No Strike	$\mathbb{E}[u_D(\min\{v_2, \mathbb{E}[v_3]\})]$, $\mathbb{E}[u_P(\min\{v_2, \mathbb{E}[v_3]\})]$	$u_D(v_1)$, $u_P(v_1)$

juror 2 when both attorneys strike. Note also that the prosecutor would be indifferent between striking and not striking only when $\mathbb{E}[u(v_2)] = \mathbb{E}[u(\max\{v_2, \mathbb{E}[v_3]\})]$, in other words when the prosecutor would choose juror 2 for any realization of π_2. The same logic holds for the defense attorney. As such, the ultimate outcome will be the same whether the attorney strikes or not, and assuming that the attorney chooses not to strike in those cases—and thereby ignoring the strike/strike box—does not change the jury selection in equilibrium.[4]

More specifically, since the attorney payoffs are 1 for winning and 0 for losing, the attorney payoffs are:

Table 1.3: Attorney Payoffs

		Prosecutor: Strike	Prosecutor: No Strike
Defense	S	$1 - c(\omega, g_2)$, $c(\omega, g_2)$	$1 - \mathbb{E}[\max(c(\omega, \pi_2), c(\omega, g_3))]$, $\mathbb{E}[\max(c(\omega, \pi_2), c(\omega, g_3))]$
	NS	$1 - \mathbb{E}[\min(c(\omega, \pi_2), c(\omega, g_3))]$, $\mathbb{E}[\min(c(\omega, \pi_2), c(\omega, g_3))]$	$1 - c(\omega, \pi_1)$, $c(\omega, \pi_1)$

So, the prosecutor will strike if

$$c(\omega, \pi_1) < \mathbb{E}\Big[\min\left\{c(\omega, \pi_2), \mathbb{E}[c(\omega, \pi_3)|g_3]\right\}\Big|g_2\Big]$$

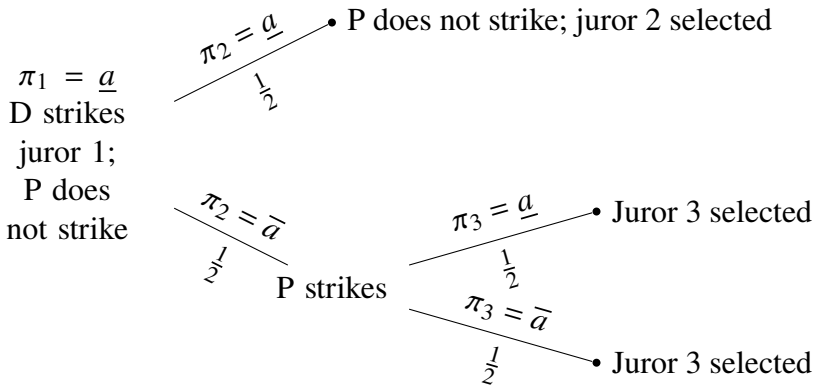
and the defense attorney will strike if

$$c(\omega, \pi_1) > \mathbb{E}\Big[\max\left\{c(\omega, \pi_2), \mathbb{E}[c(\omega, \pi_3)|g_3]\right\}\Big|g_2\Big]$$

Appendix A contains a full analysis of attorney behavior, and Appendix B derives the resulting distributions of juror types.

[4]The model assumes here that when both attorneys strike, they both lose their strikes and the juror is stricken. Note that this analysis holds for any game sequence where the payoff for an attorney of striking when the other attorney strikes is $< \mathbb{E}[u(better(v_2, \mathbb{E}[v_3]))]$. The strict inequality would make the equilibrium unique, unlike the weak inequality used here, but as explained in the text, the multiple equilibria here have the same juror selection results.

Figure 1.1: Attorney Strike Choices for Group Sequence AAA



Special Case: Zero-Variance Minority, $\underline{b} = \overline{b}$

The special case in which group B has a single value b for π_i simplifies the attorney behavior calculations because knowing that a juror is from group B fully determines the juror's cutoff s_i^* for conviction. Then, the attorney behavior determines the juror distributions (the probability that the selected juror has each of the possible payoff values: $\underline{a}, \overline{a}, b$). Derivations of the juror distributions for two of the possible sequences of groups in the jury pool illustrate how the calculations work; Appendix C contains the derivations for the remainder of the group sequences.

Group Sequence AAA

When all three jurors in the pool come from group A, an attorney who strikes juror 1 knows that his opponent will keep the second juror only if she is the worse type for him. So, striking juror 1 gives the attorney his worse type of A juror three-quarters of the time and his better type one-quarter of the time. As this expectation is better than having the worse type with certainty, juror 1 will always be struck because juror 1 will always be the worse type of juror for one of the attorneys.

So, when $\pi_1 = \underline{a}$, the selected juror will have $\pi_i = \underline{a}$ three-quarters of the time and $\pi_i = \overline{a}$ one-quarter of the time. When $\pi_1 = \overline{a}$, the selected juror will have $\pi_i = \overline{a}$ three-quarters of the time and $\pi_i = \underline{a}$ one-quarter of the time. Since $\pi_1 = \underline{a}$ with probability $\frac{1}{2}$, the juror distribution for this sequence of jury pool groups is

$\underline{a}$	$\overline{a}$	b
$\frac{1}{2}$	$\frac{1}{2}$	0

Group Sequence AAB

The juror distributions here will depend on the probability of conviction given that the juror comes from group B relative to the probabilities of conviction given that the juror has $\pi_i = \underline{a}$ and $\overline{a}$. In this special case, the probability of conviction from a group B juror equals the probability of conviction from a juror with $\pi_i = b$.

If $c(\omega, b) > c(\omega, \underline{a}) \geq c(\omega, \overline{a})$, the defense attorney will never strike the first juror because it would give the prosecutor the option to strike the second juror and have a juror with the highest known probability of conviction. Similarly, if the prosecutor strikes the first juror, the defense attorney will never strike the second juror. So, if $\pi_1 = \underline{a}$ no one will strike and if $\pi_1 = \overline{a}$ the prosecutor will strike and the defense will not, giving equal chances of the selected juror having $\pi_2 = \underline{a}$ and $\overline{a}$. These behaviors yield the following juror distribution:

$\underline{a}$	$\overline{a}$	b
$\frac{3}{4}$	$\frac{1}{4}$	0

If $c(\omega, \underline{a}) \geq c(\omega, b) > c(\omega, \overline{a})$, the defense attorney will strike juror 1 if $\pi_1 = \underline{a}$ and the prosecutor will strike juror 1 if $\pi_1 = \overline{a}$. The attorney with the remaining strike will keep juror 2 half the time (if π_2 is the better value for that attorney) and otherwise will strike juror 2.

$\underline{a}$	$\overline{a}$	b
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$

If $c(\omega, \underline{a}) \geq c(\omega, \overline{a}) \geq c(\omega, b)$, the prosecutor never strikes the first juror or the second juror, and the defense attorney only strikes when $\pi_1 = \underline{a}$.

$\underline{a}$	$\overline{a}$	b
$\frac{1}{4}$	$\frac{3}{4}$	0

Note that the first and third cases illustrate the pull of the "invisible" minority juror, who is in the pool but not seen on the eventual jury. When $c(\omega, b)$ is high, the b pool member will never be chosen because of the defense's strike. However, that

defense strategy implies that the defense cannot use a strike on the more likely to convict group A pool member, and the selected majority juror skews towards $\underline{a}$. The prosecutor mirrors these choices when $c(\omega, b)$ is low, leading to a skew towards $\overline{a}$.

All Group Sequences

Table 1.4 compiles these juror distributions, as $c(\omega, b)$ varies. It gives the overall juror distribution across all group sequences for each value of $c(\omega, b)$, assuming that jury pools are drawn randomly from the population. Finally, it notes whether the probability that the selected juror comes from group A is larger than α — in other words, whether group A is overrepresented on juries.

Table 1.4: Special Case: Juror Distributions $(\underline{a}, \overline{a}, b)$ by Group Sequence, Overall Juror Distribution, and Probability of Selected Juror Belonging to Group A as $c(\omega, b)$ Varies and with the Proportion α of the Overall Population in Group A

$c(\omega, b)$ is in the interval:	$(0, c(\omega, \overline{a})]$	$[c(\omega, \overline{a}),$ $\frac{1}{4}c(\omega, \underline{a}) + \frac{3}{4}c(\omega, \overline{a})]$	$[\frac{1}{4}c(\omega, \underline{a}) + \frac{3}{4}c(\omega, \overline{a}),$ $c(\omega, A)]$	$[c(\omega, A),$ $\frac{3}{4}c(\omega, \underline{a}) + \frac{1}{4}c(\omega, \overline{a})]$	$[\frac{3}{4}c(\omega, \underline{a}) + \frac{1}{4}c(\omega, \overline{a}),$ $c(\omega, \underline{a})]$	$[c(\omega, \underline{a}), 1)$
AAA	$\frac{1}{2}, \frac{1}{2}, 0$					
AAB	$\frac{1}{4}, \frac{3}{4}, 0$	$\frac{1}{4}, \frac{1}{4}, \frac{1}{2}$				$\frac{3}{4}, \frac{1}{4}, 0$
ABA	$\frac{1}{4}, \frac{3}{4}, 0$	$\frac{1}{4}, \frac{1}{4}, \frac{1}{2}$				$\frac{3}{4}, \frac{1}{4}, 0$
BAA	$\frac{1}{4}, \frac{3}{4}, 0$		0, 0, 1		$\frac{3}{4}, \frac{1}{4}, 0$	
ABB	0, 0, 1					
BAB	0, 0, 1					
BBA	0, 0, 1					
BBB	0, 0, 1					
$\mathbb{P}[\pi = \underline{a}]$	$\frac{1}{2}\alpha^3 + \frac{3}{4}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{3}{4}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{1}{2}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{1}{2}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{5}{4}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{9}{4}\alpha^2(1-\alpha)$
$\mathbb{P}[\pi = \overline{a}]$	$\frac{1}{2}\alpha^3 + \frac{9}{4}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{5}{4}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{1}{2}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{1}{2}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{3}{4}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{3}{4}\alpha^2(1-\alpha)$
$\mathbb{P}[\pi = b]$	$3\alpha(1-\alpha)^2 + (1-\alpha)^3$	$\alpha^2(1-\alpha)+$ $3\alpha(1-\alpha)^2 + (1-\alpha)^3$	$2\alpha^2(1-\alpha)+$ $3\alpha(1-\alpha)^2 + (1-\alpha)^3$	$2\alpha^2(1-\alpha)+$ $3\alpha(1-\alpha)^2 + (1-\alpha)^3$	$\alpha^2(1-\alpha)+$ $3\alpha(1-\alpha)^2 + (1-\alpha)^3$	$3\alpha(1-\alpha)^2 + (1-\alpha)^3$
$\mathbb{P}[g = A] > \alpha$?	Yes, if $\alpha > 0.5$	Never	Never	Never	Never	Yes, if $\alpha > 0.5$

1.4 Jury Composition, Skew of Selected Jurors, and Conviction Rate: Comparing to the Elimination of Peremptories

This section characterizes policy-relevant theoretical predictions from the zero-variance minority special case. The results are broadly divided between cases determined by the location of the minority type b in relation to the majority types:

Definition 2. *A zero-variance minority group B is* ***distinctive*** *if*

$$c(\omega, b) \notin [c(\omega, \overline{a}), c(\omega, \underline{a})]$$

and is ***intermediate*** *if*

$$c(\omega, b) \in [c(\omega, \overline{a}), c(\omega, \underline{a})]$$

Lemma 1. *A minority is distinctive when the defendant is guilty if and only if it is distinctive when the defendant is not guilty.*

This lemma follows directly from the expansion of $c(\omega, \pi_i)$ and Assumption 1 that $\nu, \gamma > 0$. Note that because $c(\omega, \pi_i)$ is not linear, $c(0, b)$ and $c(1, b)$ may both be intermediate but may exist in different sub-regions of the intermediate interval, meaning that attorney behavior and the resulting juror distribution may be different for the different states of defendant guilt.

Predictions for a Distinctive Minority

A random draw of a juror reflects a system without peremptory challenges. For a distinctive minority, the model predicts—comparing to a random draw—that the majority group will be overrepresented (and the minority underrepresented), that the selected majority jurors will be skewed towards the type that is closer to the minority type, and that the conviction rate will be closer to the minority group's conviction rate if the two majority types are sufficiently far apart when compared to the distance between the minority type and the majority type closer to it.

Proposition 1. *When the tendency to convict of a zero-variance minority is distinctive, meaning that* $c(\omega, b) \notin [c(\omega, \overline{a}), c(\omega, \underline{a})]$*,*

1. *the majority is overrepresented on the jury,*
2. *the selected majority jurors are more likely to be the type with preferences closer to the minority, and*

3. *for* $c(\omega, b) < c(\omega, \overline{a})$,

    a) *the selected jury's expected probability of conviction will be lower than the expected probability of conviction of a random draw from the population if* $c(\omega, \underline{a}) - c(\omega, \overline{a})$ *is at least four times* $c(\omega, \overline{a}) - c(\omega, b)$, *and*

    b) *as the proportion of the population in the minority increases, that ratio requirement relaxes until it becomes* $c(\omega, \underline{a}) - c(\omega, \overline{a}) > 0$ *at* $\alpha = \frac{1}{2}$

and symmetric results occur for $c(\omega, b) > c(\omega, \underline{a})$.

The first two results in Proposition 1 follow directly from Table 1.4. From Lemma 1, the result holding in the guilty state implies that it will also hold in the not guilty state. So, though the values in Table 1.4 are state-specific, the conclusions must hold in both states if they hold in one.

Note that the selected juror for all group sequences belongs to the group with more members in the pool. The attorney threatened by $\pi = b$ knows that it is the worst possible outcome for him and the best possible outcome for his opponent. When there are two A jurors, the attorney's objective will be to avoid the b juror by striking her if she appears first in the sequence and by not striking the first juror if she appears later in the sequence, since striking the first juror would give the opposing attorney the option to choose the b juror. When there are two b jurors, the attorney favoring the b juror can always guarantee the selection of a b juror. So, the proportion of selected jurors from group A equals the proportion of jury pools that have a majority of members from group A. If A makes up a majority of the population, then there is a disproportionate number of pools with a majority of members from A.

The pull of the selected A juror towards the b juror occurs because a b juror appearing second or third in the sequence threatens one attorney into accepting the type of A juror that is worse for him. In contrast, his opponent will strike the A juror that is worse for her, since she prefers the random draw from group A that results from that strike.

The third result in Proposition 1 comes from comparing the probability of conviction

for the selected jury, which is

$$\left[\frac{1}{2}\alpha^3 + \frac{3}{4}\alpha^2(1-\alpha)\right]c(\omega,\underline{a})+$$
$$\left[\frac{1}{2}\alpha^3 + \frac{9}{4}\alpha^2(1-\alpha)\right]c(\omega,\overline{a})+$$
$$\left[3\alpha(1-\alpha)^2 + (1-\alpha)^3\right]c(\omega,b)$$

for low $c(\omega, b)$, with the probability of conviction from a randomly drawn juror

$$\frac{\alpha}{2}c(\omega,\underline{a})+$$
$$\frac{\alpha}{2}c(\omega,\overline{a})+$$
$$[1-\alpha]c(\omega,b)$$

So, the defense prefers the peremptory strike system to a random draw when

$$\left[\frac{1}{2}\alpha^3 + \frac{3}{4}\alpha^2(1-\alpha) - \frac{\alpha}{2}\right]c(\omega,\underline{a})+$$
$$\left[\frac{1}{2}\alpha^3 + \frac{9}{4}\alpha^2(1-\alpha) - \frac{\alpha}{2}\right]c(\omega,\overline{a})+$$
$$\left[3\alpha(1-\alpha)^2 + (1-\alpha)^3 - (1-\alpha)\right]c(\omega,b) < 0$$

Rearranging,

$$c(\omega,\underline{a}) - c(\omega,\overline{a}) > \frac{-8\alpha^3 + 12\alpha^2 - 4\alpha}{\alpha^3 - 3\alpha^2 + 2\alpha}\left[c(\omega,\overline{a}) - c(\omega,b)\right]$$

As α increases from $\frac{1}{2}$ to 1, the right-hand side coefficient $\dfrac{-8\alpha^3 + 12\alpha^2 - 4\alpha}{\alpha^3 - 3\alpha^2 + 2\alpha}$ increases from 0 to 4. So, if the spread between the majority members is at least four times the spread between minority-leaning majority member and the minority, the defense prefers peremptory strikes, regardless of the value of α. When $\alpha = \frac{2}{3}$, the coefficient is 1 and the spread within the majority needs to be at least as large as the spread between the majority and the minority. As α decreases towards $\frac{1}{2}$, meaning that the minority is a larger proportion of the population, the condition becomes more lax.

In short, if the minority is a sufficiently large portion of the population and $c(\omega,\overline{a})$ sits close enough to $c(\omega, b)$ relative to $c(\omega,\underline{a})$, then the defense prefers peremptory strikes to a random draw of jurors. Otherwise, the prosecution will prefer peremptory strikes to a random draw.

Note that this third result of Proposition 1 applies within a particular defendant state. Due to the nonlinearity of $c(\omega, \pi_i)$, the ratio of differences $\dfrac{c(\omega, \underline{a}) - c(\omega, \overline{a})}{c(\omega, \overline{a}) - c(\omega, b)}$ may not be equal when ω changes from 0 to 1. As such, it is possible for the defense to prefer the peremptory strike process for one defendant state and the random draw for the other. Overall, the defense prefers peremptory strikes if

$$\rho\left[c(1,\underline{a}) - c(1,\overline{a})\right] + (1-\rho)\left[c(0,\underline{a}) - c(0,\overline{a})\right] > \frac{-8\alpha^3 + 12\alpha^2 - 4\alpha}{\alpha^3 - 3\alpha^2 + 2\alpha}\left(\rho\left[c(1,\overline{a}) - c(1,b)\right] + (1-\rho)\left[c(0,\overline{a}) - c(0,b)\right]\right)$$

Predictions for an Intermediate Minority

For an intermediate minority, the model predicts—again comparing to a random draw—that the majority group will be underrepresented (and the minority overrepresented), that the selected majority jurors will be skewed towards the type that is closer to the minority type only when the minority type is close enough to one of the majority types, and that the conviction rate will be closer to the minority group's conviction rate.

Proposition 2. *When the tendency to convict of a zero-variance minority is intermediate, meaning that* $c(\omega, b) \in [c(\omega, \overline{a}), c(\omega, \underline{a})]$,

1. *the majority is underrepresented on the jury,*
2. a) *when* $c(\omega, b) \in \left[\frac{1}{4}c(\omega, \underline{a}) + \frac{3}{4}c(\omega, \overline{a}), \left[\frac{3}{4}c(\omega, \underline{a}) + \frac{1}{4}c(\omega, \overline{a})\right]\right.$, *the distribution of types of majority jurors is the same as the distribution of types of majority group members,*

   b) *when* $c(\omega, b)$ *is outside that range, the selected majority jurors are more likely to be the type with preferences closer to the minority, and*
3. *the selected jury's expected probability of conviction is below (above) that of a random draw from the population if the expected probability of conviction of a minority juror is below (above) that of an average majority juror.*

The first two results in Proposition 2 appear in Table 1.4 for a particular state. Since the first result occurs in all sub-regions of the intermediate interval, by Lemma 1, it must hold in both the guilty and not guilty states if it holds in one. Note, however, that $c(\omega, b)$ may be in different subregions for $\omega = 0$ and $\omega = 1$, and so for the

same values of $\underline{a}.\overline{a}$ and b, result 2(a) may apply in one state and 2(b) in the other. However, in all subregions, the selected majority members are weakly more likely to be the type with preferences closer to the minority.

Here, because both attorneys prefer b to one of the A types, they also prefer their opponents' choice between an A and a b to that A type. So, instead of holding a strike on a first juror revealed to have the bad A type, both attorneys will strike such a juror, allowing for a b juror to be selected even when a majority of the pool comes from group A. When the majority of the pool are b jurors, both attorneys will still strike a bad A type and will not wish to give their opponents a choice between an A and a b, since that choice will be weakly worse than a b juror. So, b jurors are selected from pools with majorities from A but no A jurors are selected from majority-b pools, leading to the underrepresentation of the majority.

The pull of the b pool member on the selected A juror no longer appears in all group sequences with two A jurors. When the b juror comes second or third, the distribution of the A jurors is even. Each attorney strikes the bad A type, so the first juror is always struck. If the second juror is an A, then the opponent will keep only the struck first juror's A type. If the second juror is a b, then the opponent will either keep the b or take a random draw of an A juror. So, the potential for the minority pool member to influence the selected majority juror comes only in the BAA sequence, where an attorney will strike the first juror only if it is better to let the opponent choose between the second and third A jurors. This condition is met in the outer regions of the interval, but not in its center, since in the center the b pool member is not bad enough to make it worth giving the opposing attorney control.

Since b jurors are overrepresented and A jurors weakly favor the A type closer to b, the selected jury's expected probability of conviction will always skew towards the expected probability of conviction of a b juror. Note that this result will apply in both states.

1.5 Model Variant: Struck Jury (Attorneys Know Types)

The baseline model reflects the most commonly used jury selection procedure: strike-and-replace, where potential jurors are questioned one-at-a-time or in small groups and then strikes are decided prior to questioning the next potential juror. The other commonly used procedure in the United States is the struck jury method, where attorneys evaluate the entire jury pool at one time before making their strike decisions. This section models this struck jury procedure, both because of its

common use and because it reflects the procedure used in Mississippi that generated the data for the empirical analysis below.

Sequence: Struck Jury

The model set-up for this variant differs only in sequence, as both assume the same juror behavior and attorney preferences. As before, the attorneys each have one strike to veto a potential juror, and they know the order of the jurors. However, in the struck jury version, the attorneys know the types of all three jurors prior to making any strike decisions:

1. Nature chooses the guilt of the defendant ω and the three jurors' payoffs π_1, π_2, π_3, which fully determine the jurors' conviction thresholds s_1^*, s_2^*, s_3^*.
2. The attorneys learn ω; jurors know only the prior probability of guilt ρ.
3. **Attorneys question all jurors and learn their types** π_1, π_2, π_3.
4. Until a juror is selected to be the jury, for each juror in order, any attorneys with a remaining strike simultaneously choose whether to strike the juror, and
   - if neither attorney strikes the juror, the juror becomes the jury,
   - any attorney who uses a strike can never strike again, and
   - if an attorney strikes the juror, the process repeats for the next juror in order.
5. A trial occurs, and the selected juror obtains a signal s and votes to acquit or convict ($v = 0$ or $v = 1$), and payoffs are realized for the selected juror and the attorneys.

Attorney Behavior in Equilibrium: Struck Jury

Having known types simplifies equilibrium analysis. An attorney choosing whether to strike juror 2 will simply compare π_2 to π_3 and strike juror 2 only if π_2 is worse for him. When the attorneys decide whether to strike juror 1, they know that if they use a strike and their opponent does not, the worse of jurors 2 and 3 will become the final jury. If juror 1 has an extreme probability of conviction that is higher or lower than the other two pool members, then the attorney threatened by that juror will strike. The opposing attorney will then choose her strike so that a juror with the median probability of conviction becomes the jury. If the first juror is a median

juror, neither attorney will strike and give the opposing attorney a choice between the remaining potential jurors. So, the median juror will always be chosen, and this sequential process is equivalent to a simultaneous one in which the whole jury pool is reviewed at once.

Zero-Variance Minority, $\underline{b} = \overline{b}$: Struck Jury

When there is only one type of B juror, with $\pi_i = b$, there are 27 possible type sequences for the three-member jury pool. Note that if two jurors in the pool are of the same type, then the median, selected juror always will be of that type. The only type sequences where there are not two jurors of the same type are the six sequences that are permutations of the three types: $(\underline{a}, \overline{a}, b)$.

The probabilities of a type $\underline{a}$ juror and a type $\overline{a}$ juror are both $\frac{1}{2}\alpha$, and the probability of a type b juror is $(1 - \alpha)$. Using these probabilities to calculate the probability of each type sequence and then determining the selected juror for each type sequence yields a juror distribution with a probability of

$$\frac{1}{2}\alpha^3 + \frac{3}{4}\alpha^2(1 - \alpha)$$

for each type of juror from group A,

$$3\alpha(1 - \alpha)^2 + (1 - \alpha)^3$$

for juror type b, and an additional

$$\frac{3}{2}\alpha^2(1 - \alpha)$$

for the type of juror that has the median probability of conviction among the three types.

This additional weight to the median type of juror, means that for a distinctive minority, the probability of the A type closer to b will be increased, and the results will be identical to the strike-and-replace version of the model.

However, for an intermediate minority, the probability of the b juror will always be increased, and there will be no skew in the selected majority juror. As noted in the discussion of the strike-and-replace results, the skew for an intermediate minority in that version of the model stems from the uncertainty in unquestioned A jurors later in the order: for the BAA sequence, an intermediate b may still be worse than the expectation of their opponent's choice from two A jurors with unknown

types because of the chance that both jurors 2 and 3 might be the better type. This condition does not exist in the struck jury model because all types are known prior to the strike decisions being made.

Proposition 3. *For a zero-variance minority, when attorneys can observe the type sequence of jurors from the beginning of the jury selection process,*

1. *if the minority is distinctive, then all results are the same as for the strike-and-replace model, but*
2. *if the minority is intermediate,*
   - *the majority is underrepresented on the jury, and*
   - *the selected jury's expected probability of conviction is below (above) that of a random draw from the population if the expected probability of conviction of a minority juror is below (above) that of an average majority juror*

   as in the strike-and-replace model, but
   - *the distribution of types of majority jurors is the same as the distribution of types of majority group members.*

The zero-variance results for both the baseline model and the struck jury variant are illustrated in 1.2, 1.3, and 1.4.

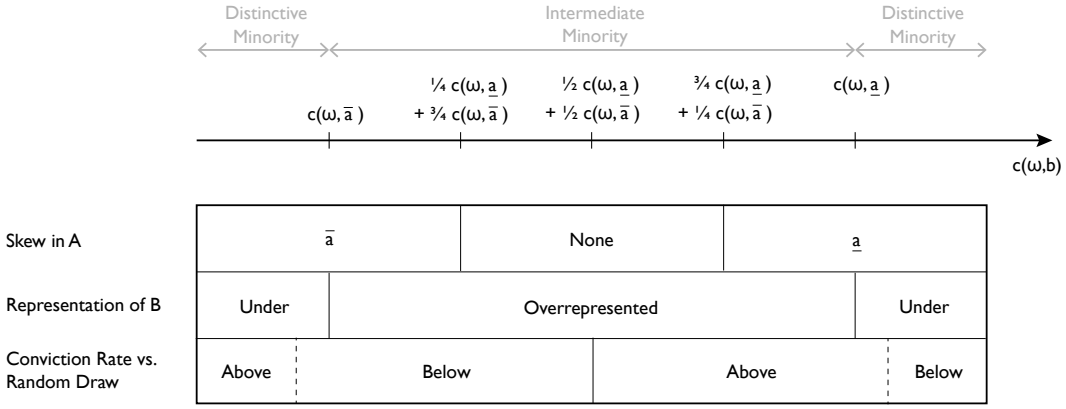


Figure 1.2: Predictions from the Zero-Variance Strike-and-Replace Model

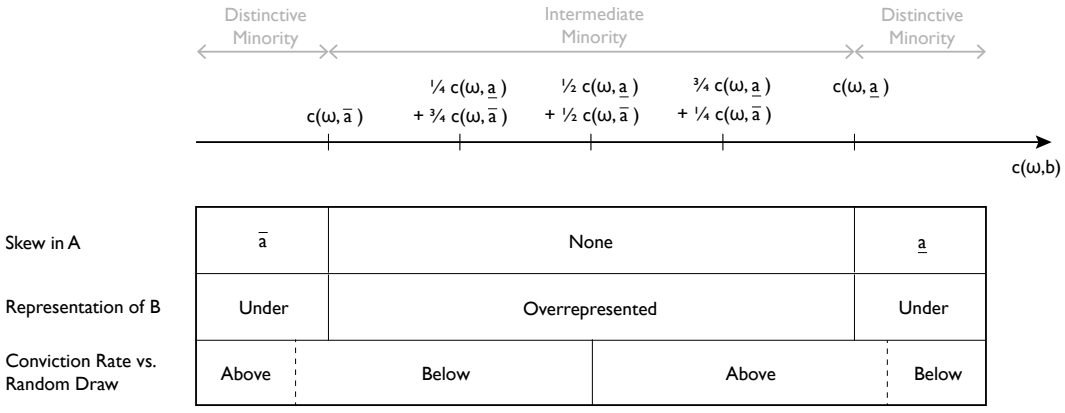


Figure 1.3: Predictions from the Zero-Variance Struck Jury Model

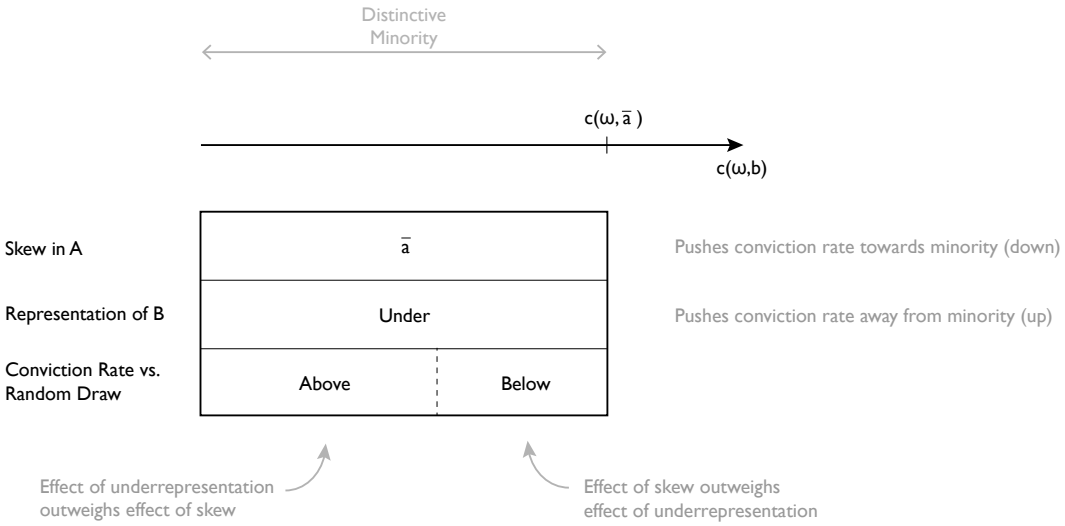


Figure 1.4: A Closer Look at the Distinctive Minority Predictions

Skew Results for Positive-Variance Minority: Struck Jury

This section generalizes the skew results for the struck jury model beyond the special zero-variance minority case. Having four types of jurors, $\pi_i \in \{\underline{a}, \overline{a}, \underline{b}, \overline{b}\}$ yields 64 possible sequences of three jurors. Again, if two jurors in the pool are of the same type, then the median, selected juror will be of that type. The only type sequences where there are not two jurors of the same type are the 24 sequences that are permutations of the four sets of three different types. There are three cases, depending on the values of the types:

1. $\underline{b}$ and $\overline{b}$ are both between $\underline{a}$ and $\overline{a}$[5]

   Since each sequence will always have at least one juror from group B, a B

[5]Or vice versa, if B is not assumed to have lower variance than A.

juror always will be selected as the jury. The flat prior over the two types of B jurors and the symmetries in the sequences—6 permutations each of $(\underline{a}, \overline{a}, \underline{b})$, $(\underline{a}, \overline{a}, \overline{b})$, $(\underline{a}, \underline{b}, \overline{b})$, $(\overline{a}, \underline{b}, \overline{b})$—mean that $\underline{b}$ and $\overline{b}$ jurors will each be selected half of the time.

2. $\underline{b}$ and $\overline{b}$ are both less than (or greater than) $\underline{a}$ and $\overline{a}$
   In the sequences with 2 A jurors, the A type closer to the B jurors always will be selected. In the sequences with 2 B jurors, the B type closer to the A jurors always will be selected.

3. $\underline{b}$ and $\overline{b}$ alternate with $\underline{a}$ and $\overline{a}$
   In each sequence, the chosen juror will be one that lies between the two types of the other group.

Note that there is no skew in the distribution of selected jurors from a group in the first case, but in both other cases, selected jurors skew towards the type closer to the mean of the *other* group.

So, for the generalized version of the struck jury model in which B also has a two-point distribution, there will be a skew in the selection of jurors from both groups, unless both types for one group are between the two types for the other group. In that case, the results will parallel the intermediate minority results in the zero-variance special case.

1.6 Data

Investigative reporters conducted an extensive review of court records in the Fifth Circuit Court District of Mississippi to produce data now available under a CC BY 4.0 license. (Craft, 2018) Unlike many jury datasets that only cover selected jurors, this data includes demographic information on jury pool members. Even more unusually, the data includes information from the questioning of potential jurors, with pool members' responses coded by the reporters in 61 binary descriptors. This combination of demographic and voir dire data facilitates empirical investigation of the theoretical predictions above.

Mississippi's Fifth Circuit uses a struck jury procedure, following Miss. R. Crim. P. 18, where the jury pool is seated in a known order. Most voir dire questions are asked of the entire pool, with follow-up questions to positive responses. After questioning ends, strikes for cause are completed. The prosecution and defense get equal numbers of peremptory challenges: 12 each when the punishment may be death or life

imprisonment, 6 each for other felony cases, and 2 each for the six-person jury for a misdemeanor. First, the prosecution considers the pool members in order, accepting or challenging them until a panel—the size of the eventual jury—has been accepted. Second, the defense may exercise peremptory challenges on the panel. These two steps repeat until there is a full panel of jurors unchallenged by either side.

The data includes records for 305 of the 418 trials conducted between 1992 and 2017 in Mississippi's Fifth Circuit, which had a population of approximately 100,000 in the 2010 census. This analysis focuses on a subset of the data: pool members with

- known gender and
- known race that is either Black or White,
- who were eligible for peremptory challenges and had a known peremptory challenge outcome and
- who were involved in a trial with only one defendant.

This subset includes 2,279 jury pool members of the original 14,874, and 80 of the 305 trials. The subset is 58% female and 69% White. The full jury pool with known gender is 57% female, the full jury pool with known race is 62% White, and the pool eligible for peremptory challenges is 64% White.

One significant difference between the subset and the full pool is that the subset excludes all acquittals. The voir dire data stems from trial transcripts, which were only preserved on the request of a party—typically due to an appeal. The state cannot appeal an acquittal.

While a dataset without this selection bias would be preferable, the theoretical results above are expected to hold conditional on guilt. In addition, acquittals make up only 11% of the trials and of the jury pool members in the full dataset.

From the available data, acquittal pool members do not appear drastically different from the pool as a whole. Demographically, acquittal pool members are similar to the full pool: 59% female and 60% White. Also, within an expansion of the subset to include multiple-defendant trials, there are 138 pool members who are neither struck nor selected as jurors or alternates because they were randomly seated later in the juror order. These pool members are more likely to be representative of the pool as a whole, and as seen in Table 1.5, they differ significantly from the full subset in

their responses only to 5 of the 61 binary descriptors, without a clear bias towards the prosecution or defense. The acquittal trials do differ in the race of the *defendant*, with 21% of all pool members being involved in trials with a White defendant but 37% of acquittal pool members having a White defendant.

Table 1.5: Differences between Full Subset and Subset Members neither Struck nor Selected with p-values ≤ 0.5

	Full (n=2486)		Not Struck (n=138)			
Question	Count	%	Count	%	χ^2	p-value
Prior jury experience	325	13.1	45	32.6	39.60	3.1e−10
Fam./friend in law enf.	505	20.3	40	29.0	5.46	0.019
Know witness	205	8.2	19	13.8	4.42	0.035
In law enforcement	52	2.1	7	5.1	4.02	0.045
Know defendant	154	6.2	15	10.9	4.00	0.046
Know victim	191	7.683	4	2.899	3.683	0.055
Fam./friend accused	315	12.671	24	17.391	2.187	0.139
Fam./friend eyewitness	1	0.040	1	0.725	1.565	0.211
Eyewitness to a crime	2	0.080	1	0.725	0.784	0.376
Prior info. on case	346	13.918	15	10.870	0.783	0.376
Victim of crime	102	4.103	8	5.797	0.560	0.454

1.7 Empirical Predictions and Findings

Proxy for Juror Type

The voir dire data reflects the information available to attorneys about pool members. Much like attorneys use this information to predict the pool members' attitudes, the following analysis uses the data to construct a proxy for type. The count-based proxy in this section considers all voir dire questions with a clear pro-prosecution or pro-defense lean and simply subtracts the number of pro-defense responses from the number of pro-prosecution responses for each pool member. The categorization of the 37 questions that had any True responses as pro-prosecution, pro-defense, or ambiguous appears in Tables 1.6, 1.7, and 1.8. As seen in Appendix F, another proxy of pro-prosecution lean based on an item response theory model yields similar results, providing evidence of their robustness.

Table 1.6: Pro-Prosecution Questions

Description	**# True**
The juror has friends or family who work or have worked in law enforcement	469
Juror has friends or family that have been the victims of a crime	116
Juror has been the victim of a crime	94
The juror works or has worked in law enforcement	47
Juror expressed a bias in favor of the prosecution	15
Juror was a witness for the state in a criminal case	9
The juror served in the military	7
The juror was an eyewitness to a crime	2
Juror expressed that he or she was more likely to believe the testimony of the police over other witnesses	2

Table 1.7: Pro-Defense Questions

Description	**# True**
The juror was accused of being involved in criminal activity	33
Juror expressed reservations about imposing the death penalty either because of moral, religious or ethical reasons	17
Juror expressed a bias in favor of the defense	9
Juror said he or she could not or would not impose the death penalty	2
Juror admitted he or she believes the defendant is innocent	1

Table 1.8: Ambiguous Questions

Description	**# True**
Juror has served on a jury before	309
Juror had prior information on the case	301
The juror has friends or family accused of being involved in criminal activity	292
Juror has prior familiarity with witnesses through either personal or professional channels	197
Juror has prior familiarity with attorneys or the judge through personal or professional channels	171
Juror has prior familiarity with victim through either personal or professional channels	166
Juror has prior familiarity with defendant through either personal or professional channels	139
Juror expressed a bias but not clear if it favors the state or the defense	16
Juror said his or her occupation would make serving difficult	13
Juror had medical problems preventing the juror from serving	11
Juror is unemployed	6
Juror had difficulty communicating or understanding	5
Juror was a witness in a criminal case but not specific about which side	5
Juror admitted another reason the juror would not be able to be fair	4
Juror had caretaker obligations	3
Juror was a defendant in a civil dispute	2
Juror expressed a moral or emotional hardship	2
Juror said prior social obligations would make serving difficult	2
Juror admitted to moral/religious/conscientious beliefs that would affect his or her decision or prevent them from sitting in judgment	2
Juror said he or she would have difficulty making decisions based only on evidence	2
Juror was a witness in a civil dispute	1

Examining the proxy values by race indicates that White pool members tend to be more pro-prosecution than Black pool members.

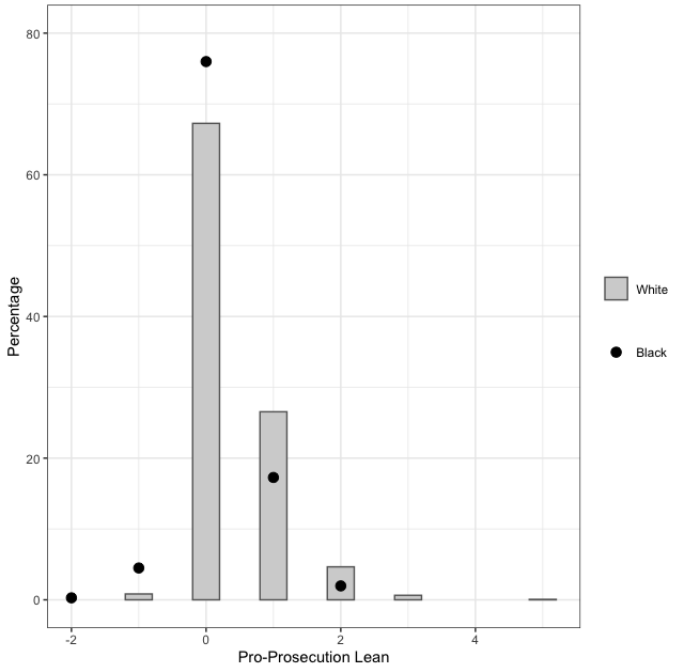


Figure 1.5: Plot of Count Proxy by Race

Examining the strikes by the prosecution and defense suggests that the proxy is aligned with attorneys' impressions from the voir dire questioning: the prosecution is more likely to strike pool members with lower values of the proxy, while the defense is more likely to strike members with higher values.

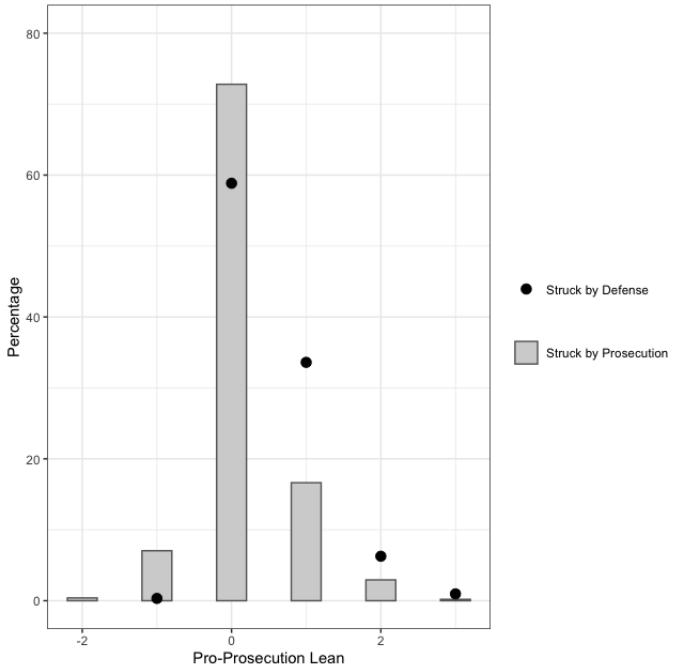


Figure 1.6: Plot of Count Proxy by Strikes

Unrepresentative Representation

Although the proxy may not be precise enough to draw clear conclusions about the shapes of the distributions for both groups, there is substantial overlap between the groups in their voir dire responses, but with a pro-defense tail composed of Black pool members and a pro-prosecution tail composed of White pool members. This situation is closest to a version of the struck jury model where both groups have two-point distributions that overlap: $\overline{b} > \overline{a} > \underline{b} > \underline{a}$. That version of the model predicts a skew towards the opposite group for the selected jurors from both the majority and the minority.

The proxy exhibits a significant pro-defense skew in the selected White jurors. The pro-prosecution skew in selected Black jurors exists as well, but is not significant in the current data. These skews can be seen in the raw data, as well as in the regressions of

$$Proxy_i \sim \beta_0 + \beta_1 White_i + \beta_2 White_i * Selected_i + \beta_3 Black_i * Selected_i$$

and the corresponding interaction plot. The pro-prosecution lean is reduced by 0.186 for selected White jurors as compared to unselected White pool members.

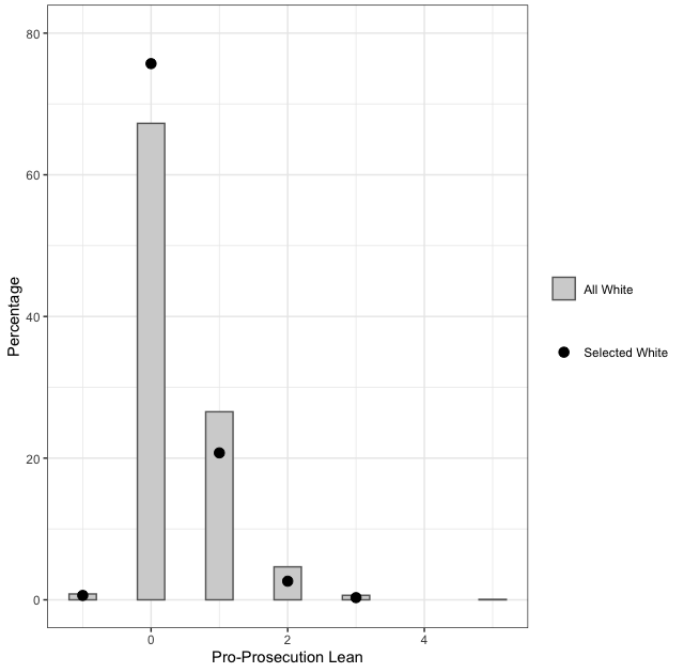


Figure 1.7: Plot of Count Proxy by Selected and Unselected White Jurors

Table 1.9: Count-based Proxy Regression Results

	Dependent variable:
	Pro-Prosecution Lean
White	0.305*** (0.034)
White * Selected Juror	−0.186*** (0.031)
Black * Selected Juror	0.046 (0.046)
Constant	0.144*** (0.028)
Observations	2,279
R^2	0.042
Adjusted R^2	0.041
Residual Std. Error	0.595 (df = 2275)
F Statistic	33.312*** (df = 3; 2275)
Note:	*p<0.1; **p<0.05; ***p<0.01

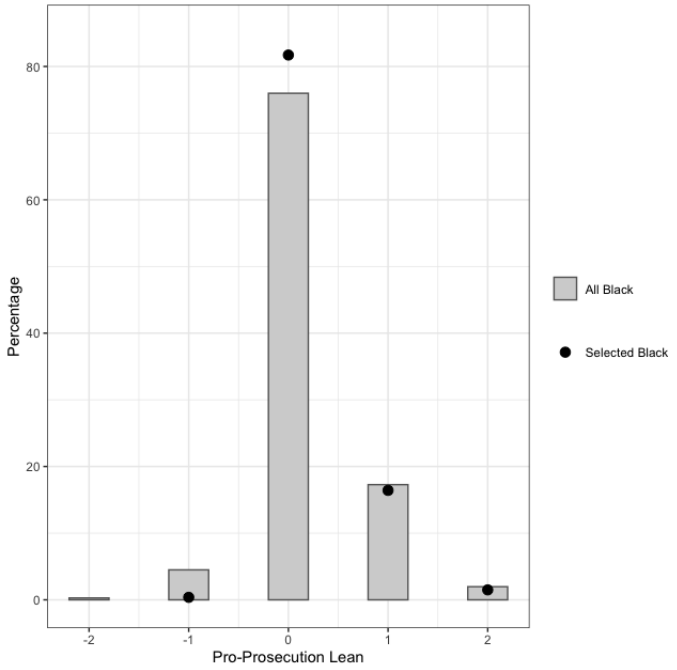


Figure 1.8: Plot of Count Proxy by Selected and Unselected Black Jurors

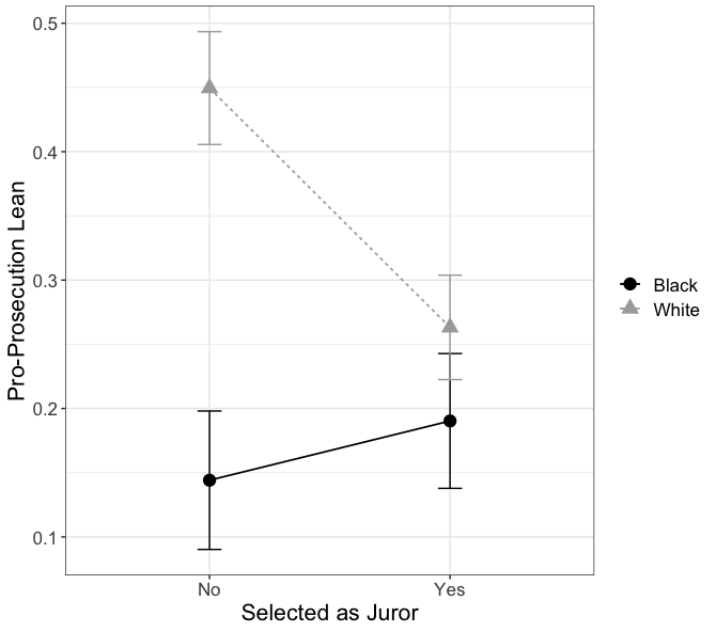


Figure 1.9: Interaction Plot of Count Proxy

1.8 Conclusion

This paper examines, both theoretically and empirically, the peremptory challenge process in jury selection. Theoretically, it fits into a small literature of game theoretic models of peremptories in which each potential juror has a type, representing his pro-prosecution lean. It complicates the typical model set-up by separating the jury pool into two groups, each with its own distribution of types. These choices of distributions are novel and yield new predictions about the effects of peremptory challenges on minority representation and conviction rates. Empirically, it takes advantage of a little-used dataset with unusually comprehensive information about potential jurors. It uses data from the voir dire questioning to create a proxy for potential jurors' types that can then help test the predictions from the theoretical model.

The first key finding is that people selected to serve on juries can be unrepresentative of their racial groups. This unrepresentative representation occurs unless all types from the minority group are intermediate, with no type being the least or most likely to convict in the potential juror population. The process skews selected jurors away from the mean type of their own group, pushing them towards the mean type of the *other* group. The empirical results confirm that this skew happens in the proxy values for the selected jurors, offering the first direct confirmation of a behavior hypothesized in earlier works, such as Anwar et al. (2012) and Schwartz & Schwartz (1996).

Second, although prior theoretical work only predicts underrepresentation of minority groups through peremptory challenges, this model also predicts overrepresentation, again only when the types from the minority group are intermediate. The model offers a theoretical framework for prior *empirical* work finding increases in minority representation after peremptories, as seen in Rose (1999), Anwar et al. (2012), Gau (2016), and Flanagan (2018).

Third and finally, although peremptory challenges are often assumed to favor the prosecution, this model predicts decreases in the conviction rate when a pro-defense minority has sufficiently moderate types. When minority types are intermediate, this decreased conviction rate comes alongside overrepresentation of a pro-defense minority. However, when the pro-defense minority type is distinctive—meaning it has the lowest probability of conviction of all types in the potential juror population—the conviction rate may still decrease because the pro-defense skew of the selected White jurors outweighs the effect of underrepresentation of Black jurors. The

resulting jury looks more likely to convict than it is.

These results suggest that policies eliminating or modifying peremptory challenges may have qualitatively different effects on both representation and conviction rates, depending on the type distributions of the groups considered. It seems likely that these distributions will vary by geographic location and also by case characteristics, such as the defendant race or the charges being tried. More generally, policymakers should be cautious about conflating a jury matching its community's demographics with a jury matching its community's desire to convict a defendant.

Theoretically, a fruitful areas for future work include generalizations of the model, considering additional type distributions as well as expansions of the size of the jury to connect this work to the extensive literature on jury deliberation and the effects of requiring unanimity. Empirically, the construction and analysis of larger datasets with more trials would allow for comparisons of subsets of data where the type distributions are different, enabling further testing of the model predictions.

Chapter 2

ARE TWO HEADS BETTER THAN ONE? STRATEGIC IMPLICATIONS OF SPLITTING CONVICTION AND SENTENCING BETWEEN THE JURY AND JUDGE

> [T]he Guidelines . . . have made charlatans and dissemblers of us all. We spend our time plotting and scheming, bending and twisting, distorting and ignoring *the law* in an effort to achieve a just result. All under the banner of "truth in sentencing"!
> — Survey response of anonymous trial judge (Weinstein, 1992)

Criminal trial outcomes involve two steps: the determination of guilt, followed by the determination of a sentence if guilty. In U.S. noncapital cases, the decisionmakers for those steps vary across trials. Typically, a jury decides guilt, then a judge sentences. However, in some cases a single actor performs both steps: the judge in a bench trial, and the jury in the few jurisdictions permitting jury sentencing.[1]

Legal scholars have long debated the involvement of jurors in the process, due to a general consensus that "[j]urors doubtless are somewhat more lawless than judges, because they do not internalize the values of law-following to the same extent as most judges" (Posner, 1999). They worry that this lower level of lawfulness will lead to nullification (acquitting a defendant for reasons other than probability of guilt, such as the harshness of the expected sentence) and compromise verdicts (agreeing to convict someone on a lesser charge — or with a lesser punishment if the jury also controls sentencing — because of uncertainty about guilt). Jury sentencing heightens these concerns because it gives the jury control over both the verdict and the sentence.

This paper constructs a model that examines strategic decisionmaking about the verdict and sentence, within a single-actor process and a two-actor one. The baseline model involves actors with a typical, state-dependent outcome utility: an actor has an ideal sentence for someone known to be guilty and suffers losses from erroneous

[1] Though juries in the United States regularly decide on civil damages and the death penalty, jury sentencing in noncapital criminal cases has long fallen out of favor. It remains today in six states, concentrated in the South: Arkansas, Kentucky, Missouri, Oklahoma, Texas, and Virginia. (Holtzman, 2021)

verdicts and from sentences for guilty defendants that stray from the actor's ideal. In the two-actor process, the actor deciding the verdict has full information about the beliefs and preferences of the actor choosing the sentence, who cannot commit to a particular sentence. The baseline model is then extended in two ways. First, a mandatory minimum sentence restricts the discretion in the second step. Second, the actors' utility is altered to add a lawfulness utility. A lawful probability threshold for guilt and sentence for the guilty are set exogenously. The actors now suffer additional losses when they choose a verdict different from the one implied by the law's threshold and when they choose a sentence that deviates from the lawful sentence.

The results from the baseline model find that a single actor with full control over both steps always will convict the defendant and will give a punishment scaled by the probability of guilt of the defendant, learned during the trial process. In other words, the actor only acquits (or convicts and gives zero punishment to) defendants who have zero probability of guilt. A defendant with a low probability of guilt would still be convicted, but given a slap on the wrist. This result offers an extreme version of the probability-based sentencing sometimes suggested as a fair and efficient reform (Abramowicz, 2001; Fisher, 2011; Schuman, 2015; Teichman, 2017; Siegel & Strulovici, 2019). Because the probability threshold above which the actor would convict depends on the expected sentence — which in turn scales with the probability of guilt, the moving threshold always remains beneath the defendant's probability of guilt. Essentially, every case leads to a compromise verdict.

The two-actor baseline model finds that the first actor will convict when her ideal sentence is greater than a proportion of the second actor's ideal sentence. If the second actor's ideal is large enough, the first actor prefers acquittal to conviction with such a harsh sentence. These acquittals reflect jury nullification due to the expected sentence, which appears in the empirical literature. (Hannaford-Agor & Hans, 2003; Garvey et al., 2004; Bindler & Hjalmarsson, 2018) The convictions continue to be compromise verdicts, as in the single-actor case, but with the expected sentence determined by the second actor. Because both actors' sentences would scale with the probability of guilt, the decision to acquit depends only on the sentencer's ideal, and not on any specifics of the case.

The model extension with a mandatory minimum again finds a nullification-like result, where the first-step actor acquits due to the expected sentence. Here, however, because the mandatory minimum is fixed and does not scale with the probability of

guilt, the acquittals occur when the probability of guilt is low compared to a ratio of the minimum and the ideal sentence. This model version reflects mandatory minimums required by law, and it also serves as a version of the model where any conviction is costly to the defendant, even if the sentencing step results in no fine or incarceration.

The second, and main, model extension incorporates the lawfulness utility for all actors. The novel two-part utility that results reflects the idea that the actors in the process care about following the law, but also have their own innate sense of justice. As Posner (1999) notes, the "jury system presupposes some degree of compliance by jurors with the rules laid down by the judge to guide them, but not 100 percent." And, as illustrated in the epigraph above, judges also may experience tension between their desires to follow the law and their own opinions about just punishment. In fact, Bushway et al. (2012) study judges in Maryland — under voluntary sentencing guidelines — who had made accidental mistakes in calculating the guidelines recommendations, and finds that for "each additional month errantly included in the recommendation, criminal sentences increase by an average of 4 days" but "each month subtracted from a recommendation reduces sentences by an average of 13 days." They further note: "In this institutional setting characterized by substantial discretion of downstream actors, some may be surprised that these voluntary guidelines alter sentencing outcomes at all. For those whose prior is that public servants comply with directives, the surprise will be the extent of noncompliance." Considering a two-part utility reflective of these behaviors yields three main results.

First, because actors pay the lawfulness cost only for their own actions, choosing a two-actor process can create a free-riding effect. Nullification-like acquittals still occur as in the baseline model, when the sentencer and the lawful sentence are relatively harsh. However, the new free-riding outcome occurs when the first actor prefers to be more lenient than the lawful sentence and also knows that the sentencer shares that preference for leniency. An actor who would have chosen a nullification acquittal if she controlled both steps of the process now will convict because she knows the sentencer will give a reduced sentence and she will not have to pay the lawfulness cost for that deviation.

Second, in a single-actor process, increasing the lawfulness of the actor can cause non-monotonic effects on the actor's verdict choice. When an actor is completely lawless, paying zero lawfulness utility costs, the actor will convict and give their

own ideal sentence, as in the baseline model. When an actor is infinitely lawful, the actor will follow the law exactly. So, when the law favors conviction, the infinitely lawful actor will convict and give the lawful sentence. However, consider a situation where the law favors conviction and the sentencing lawfulness cost is relatively large compared to the verdict lawfulness cost. For example, the lawful sentence is significantly harsher than the actor's ideal sentence. Then, actors with an intermediate level of lawfulness will choose a nullification-like acquittal. Increasing an actor's lawfulness from zero will first yield compromise verdicts close to the actor's ideal, then nullification-like acquittals, and then compromise verdicts approaching the lawful sentence.

Third, another non-monotonic result in the opposite direction can occur, when increasing the sentence preferred by the law or a second-actor sentencer or when increasing the lawfulness of a second-actor sentencer. Here, the increase can result in acquittals, followed by compromise verdicts around the first-actor's baseline sentence, and then another region of acquittals. This effect requires that increasing the parameter causes the chosen sentence to change from being lower (higher) than the first actor's baseline sentence to higher (lower) than it. As the expected sentence approaches the first actor's baseline sentence, the first actor gains utility from conviction, but then loses utility once the sentence becomes too large.

2.1 Related Literature

This model has two distinguishing features. First, it directly considers the interplay between two actors within the process of conviction and sentencing, allowing for a comparison with the single-actor process and consideration of jury nullification on a case-specific basis rather than only at the level of setting overall sentencing policy. Second, it incorporates a two-part utility that considers lawfulness as well as outcome utility, separating the effects of these two motivations. As such, the work relates to several longstanding debates about decisionmaking in criminal trials.

Nullification and Compromise Verdicts

Jury nullification is "both prohibited and protected in a unique way," where jurors cannot be instructed about their power to nullify and potential jurors can be struck for advocating nullification, but there are essentially no ways to punish jurors for nullification or reverse an acquittal. (Duvall, 2012) Numerous scholars have argued that nullification should be encouraged, potentially by enshrining it as a legal right of the jury. In one influential article, Butler (1995) advocates for black jurors to nullify

certain cases brought against black defendants to base decisions about incarceration "on the costs and benefits to their community, than by the traditional criminal justice process, which is controlled by white lawmakers and white law enforcers." More recently, Salib & Krishnamurthi (2022) argue that "nullification may have an important role to play in blunting the force of the most extreme anti-abortion laws" as the threat of nullification may prevent prosecution in extreme cases; they suggest that a similar fear of nullification caused marijuana prosecutions to drop disproportionally compared to other drug prosecutions during a period when public support for criminalizing marijuana reduced from 63% to 32%.

Empirical evidence supports the anecdotal evidence that nullification can occur in response to excessive punishment. Bindler & Hjalmarsson (2018) use historical data from 18th and 19th century London to find that abolishing capital punishment and halting penal transportation both increased juries' likelihood of conviction, implying that the harsher initial punishments led to nullification. Garvey et al. (2004) find mixed results, where jurors' beliefs about the fairness of a criminal law or punishment affect verdicts only in certain jurisdictions and under certain circumstances. Hannaford-Agor & Hans (2003) evaluate state criminal trial data, including accompanying juror survey responses, to conclude that "it is clear . . . that juror concerns about legal fairness and outcome fairness are present to a measurable extent in hung and acquittal juries." However, Silveira (2017) uses a nonparametric estimation procedure on North Carolina data and finds that a counterfactual decrease in potential trial sentences would decrease the conviction rate at trial, though it would increase the proportion of defendants who settle through plea bargains enough that the overall proportion of punished defendants increases.[2]

Compromise verdicts and the burden of proof for conviction also have generated policy disagreement. The traditional legal standpoint holds that compromise verdicts are problematic and that the beyond-a-reasonable-doubt threshold for guilt should remain constant in all cases: "[W]e do not adjust the criminal burden of proof based on likely sanctions, although sanctions largely determine the costs paid by convicted defendants" (Lempert, 2001). However, some scholars advocate for compromise verdicts and the accompanying variable burdens of proof. For example, Schuman (2015) proposes that "drug sentences could be made shorter, fairer, and more efficient by varying the punishment imposed based on the probability that the offender trafficked a particular quantity of drugs." Hoffman (2003) — written by a Colorado

[2] Andreoni (1991) also reviews earlier literature with evidence that higher penalties decrease likelihood of conviction.

district judge — and Holtzman (2021) also offer arguments in favor of compromise verdicts, suggesting that they allow for more precision in assessments of culpability.[3]

Teichman (2017) "argues that the legal system routinely relaxes the burden of proof in criminal adjudication by adjusting the substantive content of criminal law," meaning that the ability to convict defendants with different, lesser offenses is effectively used to reduce the sentences and conviction threshold. Compromise verdicts are built into these decisions about reducing charge severity for weaker cases. Among other empirical, survey, experimental, and anecdotal evidence supporting the idea that the burden of proof is lower for smaller punishments, Guttel & Teichman (2012) describe a "strand of cases . . . in which trial judges encourage hung juries to reach a guilty verdict by assuring them that they will treat the defendant with leniency, or by allowing the jury to recommend leniency."

Existing game theoretical models considering nullification and compromise verdicts have mostly taken the sentence as exogenous, considering the welfare effects of an external policy change altering a required sentence. Andreoni (1991) argues that increasing a sentence will reduce the probability of conviction leading to the optimality of sentences that increase with the severity of a crime, in contrast to the line of argument in Becker (1968) and related works advocating for maximal sentences with varying thresholds for conviction. In this model, the juror sets for himself the threshold of reasonable doubt based on outcome utility with a fixed sentence. Fisher (2011) argues against the bifurcation of verdict and sentence, finding that creating a range of convictions (and corresponding sentences) based on probability of guilt will enhance deterrence under an assumption that defendants are risk-loving. Siegel & Strulovici (2019) find welfare improvements from adding similar intermediate verdicts.

It appears that the only model of compromise verdicts that includes sentencing discretion is Lundberg (2016b), which considers a single-actor process with a generalized outcome-based, state-dependent utility.[4] Lundberg concludes that any actor — jury or judge — who determines the verdict and also controls the sentencing will

[3] Siegel & Strulovici (2019) review additional literature that advocates for probabilistic sentencing in criminal cases as well as probability-based damages in civil cases.

[4] The theory literature on sentencing discretion instead focuses on the interplay between a higher authority and a judge, where the higher authority can give varying levels of discretion to the judge. Reinganum (2000) considers a game between a judge and a sentencing commission that decides whether to have guidelines or judicial discretion in a system with plea bargaining. Although the judge and commission have the same preferences, the judge has more information at the time of decision and the commission is also concerned about how trial outcomes affect plea bargain settlements. There are no compromise verdicts in this model because all defendants are assumed to

engage in compromise verdict-like behavior, reducing the threshold probability for conviction along with the length of the sentence. Lundberg also asserts that this tendency to compromise argues for the jury-judge two-actor process, though the "success of [that] trial format...in stopping compromise verdicts is conditional on the requirement that jurors ignore the consequence of their verdict" because "if the jury and judge have similar preferences and information, the jury will anticipate the sentence the judge will impose, and the outcome will be exactly the same as if the jury had been given discretion." However, the article does not further support the idea that jurors will ignore judges' expected sentences or model what occurs when judge preferences vary. In a follow-up empirical study, Lundberg (2016a) finds that 1) judges convict less often for crimes with higher possible sentences (though juries do not), but that 2) judges did not respond to greater sentencing discretion either by increased convictions or reduced sentences.

Characteristics of Judges and Juries

Policy arguments about jury involvement in the criminal trial process often revolve around perceived differences between juries and judges. The discussion around the practice of jury sentencing highlights these concerns. Twentieth-century scholars and practitioners almost uniformly opposed jury sentencing in noncapital cases.[5]Opponents argue that juries are more prejudiced, less uniform, and harsher than judges, and they warn about the danger of compromise verdicts.[6] Webster (1960), calling for the abolition of jury sentencing in Texas, states that "[m]ost people would be shocked at the idea that, if half of the jury found a man guilty and half found him innocent, they would then try to find some lesser crime that was acceptable to all, but this practice is often utilized, in effect, when the jury engages in fixing a term of years for a particular offense."

---

be guilty. Shavell (2007) considers when to give discretion to a judge who is more informed but also has different preferences over the sentence, modeling varying levels of discretion and controls on discretion including penalties that vary with the sentence chosen and the threat of appeals. Miceli (2008) involves a game between a legislature setting sentencing ranges and a judge choosing sentences, where the judge wants to choose a fair sentence given information about the case, and the legislature wants to choose a fair sentence but also to deter crime. The legislature sets a range around the deterrent ideal that narrows as the legislature places more weight on deterrence. The judge preferences in Shavell (2007) and Miceli (2008) are not micro-founded.

Empirically, Shepherd (2002) finds that the increased sentences from truth-in-sentencing laws cause a reduction in the crimes involved but also substitution into other crimes.

[5]Lanni (1999) notes that she was "aware of only one article, written in 1918, that supports the institution." A 1967 Presidential Commission and the American Bar Association both advocate for the abolition of jury sentencing (President's Commission on Law Enforcement and Administration of Justice (1967), American Bar Association (1994)).

[6]Hoffman (2003), among others, summarizes these arguments.

However, around the year 2000, a contingent of legal scholars began to voice support for jury sentencing, likely due to 1) dissatisfaction with the 1987 Federal Sentencing Guidelines and other mandatory and determinate sentencing regimes seen as draconian (Lanni (1999)) and 2) the greater role for the jury in determining facts related to sentencing in the Apprendi v. New Jersey (2000) line of cases. Proponents argue that juries offer a more democratic and community-oriented way to determine sentences, which is especially valuable when society lacks a clear consensus on the purpose of punishment (Iontcheva (2003)).[7]

The empirical evidence about the traits of juries and judges is mixed. Eisenberg et al. (2005) conduct a partial replication of the classic study in Kalven & Zeisel (1966), which surveyed judges in thousands of state criminal cases, finding that judges mostly agree on verdicts with juries but that disagreement mostly stems from juries acquitting where judges would have convicted. Eisenberg et al. (2005) extends this study by also surveying juries about their views on the evidentiary strength in the cases, finding again that juries overall tend to be more lenient than judges: judges convict more often in cases with moderate evidentiary strength; however, judges also sometimes disagree with juries about the strength of the evidence in both directions and causing disagreement about the verdict in both directions as well. The data also suggests that there may be variation by jurisdiction in the relative leniency of juries, with juries being more lenient in the Bronx, D.C., and Los Angeles but not in Maricopa County, though fewer than 100 trials are included for each jurisdiction.

Indeed, as Leipold (2005) notes, federal criminal defendants and defense counsel overwhelmingly prefer jury trials to bench trials. However, Leipold continues that "it is unclear why" given that bench trials have much lower conviction rates than jury trials (55% versus 84%, with bench trial conviction rates varying from 37% to 78% between the different federal circuits while jury conviction rates only vary from 80% to 89%), even when controlling for selection into bench trials based on type of crime, and "while the conviction rate for juries has remained nearly constant

[7] In the context of capital cases, Justice Breyer's concurrence in Ring v. Arizona (2002) notes:

> In respect to retribution, jurors possess an important comparative advantage over judges. In principle, they are more attuned to "the community's moral sensibility" because they "reflect more accurately the composition and experiences of the community as a whole." Hence, they are more likely to "express the conscience of the community on the ultimate question of life or death" and better able to determine in the particular case the need for retribution, namely, "an expression of the community's belief that certain crimes are themselves so grievous an affront to humanity that the only adequate response may be the penalty of death."

Jury sentencing supporters extend this reasoning to noncapital cases as well.

for many years, the judicial rate has fallen steadily since the late 1980s." Gay et al. (1989) offers one theoretical explanation for the disparity: if juries and judges are equally harsh but juries are noisier at determining guilt and are not strategic, innocent defendants will select bench trials.[8] However, this theory does not explain the decrease in bench trial convictions over time, and Leipold (2005) suggests that judges' disapproval of the harsher sentences under Federal Sentencing Guidelines created this trend, though it is unclear why defense attorneys have not changed their behavior to select bench trials more often.

Not only do judges acquit more often when they take over the jury's traditional first step in the process, but also juries sentence more harshly and with greater variance when they take over the judge's second step. King & Noble (2005) finds that jury sentencing leads to more variable and harsher sentences in Arkansas and Virginia, and Weninger (1994) finds similar results in Texas. However, it is unclear how much of this difference is due to institutional design choices aimed at keeping jury sentences high to deter defendants from choosing costly jury trials.[9] Carrington (2011) criticizes states with mandatory or "pseudomandatory" jury sentencing, where opting for a jury trial requires that the defendant also acquiesce to jury sentencing. McCloy (2021) and Klein (2021) note that Virginia juries are given information about statutory minimums and maximums but are prevented from learning about state sentencing guidelines that can recommend sentences much closer to the minimum.

Weninger (1994) also notes that public opinion surveys suggest that the Texas public may simply be harder on crime than the judiciary. More recently, Rappaport (2020) argues that democratization reforms through greater jury involvement will increase conviction rates, in opposition to the reformers' expectations. Rappaport bases this argument on an extensive review of literature "drawn from political science,

[8] An alternative theoretical explanation that defendants choose bench trials when the judge is known to be more lenient appears to be unexplored by the theoretical and empirical literature.

[9]

> The risk of a severe jury sentence is perceived by defendants to be so daunting, that prosecutors in at least one urban jurisdiction in Kentucky are able to use the threat of a jury sentence to negotiate settlements after guilty verdicts, settlements in which the defendant gives up his right to challenge his conviction or sentence on appeal in exchange for the prosecutor's sentencing recommendation and consent to waive jury sentencing. ...[T]he importance of high and variable jury sentences for encouraging jury waivers was also noted by those interviewed in Virginia, Kentucky, and Arkansas as a major roadblock in the way of abandoning jury sentencing

(King, 2004).

psychology, sociology, economics, criminology, and empirical legal studies — about public attitudes toward punishment, racial bias, judicial behavior, group decision-making, and more." About public opinion, Rappaport explains:

> First, while public opinion is certainly less punitive today than it was three decades ago, at the tail end of a massive crime wave, it remains quite harsh. A majority of the country continues to support the death penalty and still believes that courts are too lenient. Well under 20 percent of Americans think that prison conditions are too harsh. And the emergence of seemingly lenient attitudes has not crowded out punitive beliefs. In fact, numerous studies show that the same individuals who support the death penalty and view the courts as too lenient also support rehabilitation and alternative sentences to incarceration.

Other Related Literature

Other theoretical work in the judicial literature also have included multi-part utilities to capture multiple motivations of actors. Shadmehr et al. (2022) assess coordination between judges by giving them payoffs that depend on both how close their decision is to the legally correct one and how close it is to another judge's decision. Carrubba & Clark (2012) and Parameswaran et al. (2021) considers judges with preferences that have an expressive component based on a judge's individual vote, as well as a policy component. Polinsky & Shavell (2000) considers a fairness-related utility from others' punishments in addition to utility from deterrence, leading to the optimal sentence lying between the one that maximizes fairness and the one that maximizes deterrence.

2.2 Baseline Model: Only Outcome Utility

Punishment in a criminal trial is determined in two steps: first, a verdict $v \in \{0, 1\}$ determines whether the defendant is innocent or guilty; second, if found guilty, the defendant receives a sentence $s \in \mathbb{R}_0^+$. A single actor may perform both steps, or a different actor may perform each step.

Actor Utility

An actor J_i has an ideal sentence of 0 for an innocent defendant and of $s_i \in \mathbb{R}_0^+$ for a guilty defendant, with quadratic loss from these ideal sentences. So, J_i has an

outcome utility stemming from the following state-dependent payoffs

	$\omega = 0$	$\omega = 1$
$v = 0$	0	$-s_i^2$
$v = 1$	$-s^2$	$-(s - s_i)^2$

where $\omega \in \{0, 1\}$ is the defendant's innocence or guilt.

Single Actor: Sequence

First, consider the single-actor case in which the same J_i determines the verdict and the sentence.

1. Nature chooses guilt of defendant in current case $\omega \in \{0, 1\}$
2. Single actor J_i learns defendant's probability of guilt π through observation of trial
3. J_i determines verdict $v \in \{0, 1\}$ and, if guilty, determines sentence $s \in \mathbb{R}_0^+$
4. Payoff realized

Single Actor: Sentencing (Step 2)

Upon reaching the sentencing step, J_i will choose s^* to maximize expected utility:

$$\max_s (1 - \pi)(-s^2) + (\pi)(-(s - s_i)^2)$$

Then,

$$s^* = \pi s_i$$
$$EU_{v=1} = -\pi(1 - \pi)s_i^2$$

Single Actor: Verdict (Step 1)

Given the sentence that J_i would choose upon a guilty verdict, J_i will only convict if $EU_{v=1} - EU_{v=0} > 0$. In general,

$$\begin{aligned} EU_{v=0} &= (1 - \pi)(0) + \pi(-s_i^2) = -\pi s_i^2 \\ EU_{v=1,s} - EU_{v=0} &= (1 - \pi)(-s^2) + (\pi)(-(s - s_i)^2) + \pi s_i^2 \\ &= 2\pi s s_i - s^2 \end{aligned}$$

and for $s^* = \pi s_i$,

$$EU_{v=1,s=\pi s_i} - EU_{v=0} = \pi^2 s_i^2$$

Thus, J_i is indifferent between conviction and acquittal when the probability of guilt $\pi = 0$, but will convict and sentence πs_i for all $\pi > 0$. Because J_i can control the sentence precisely, they are willing to convict anyone with any positive probability of guilt, giving minor sentences for low probabilities of guilt. This behavior is an extreme version of a compromise verdict.

Two Actors: Sequence

The sequence remains similar for the two-actor case:

1. Nature chooses guilt of defendant in current case $\omega \in \{0, 1\}$
2. Two actors J_i with $i \in \{1, 2\}$ learn probability of guilt for specific case π through observation of trial
3. J_1 determines verdict $v \in \{0, 1\}$ with full knowledge of J_2's payoff structure
4. If verdict is guilty, J_2 determines sentence $s \in \mathbb{R}_0^+$
5. Payoffs realized

Two Actors: Verdict (Step 1)

The second actor J_2 will select a sentence exactly as in the single-actor case, and so $s^* = \pi s_2$. Then, the first actor J_1 will convict if

$$\begin{aligned} EU_{v=1,s=\pi s_2} - EU_{v=0} &> 0 \\ 2\pi^2 s_1 s_2 - \pi^2 s_2^2 &> 0 \end{aligned}$$

meaning that when $\pi = 0$, J_1 will be indifferent between conviction and acquittal since $s^* = 0$ and that when $\pi > 0$, J_1 will convict when $s_1 > \frac{s_2}{2}$. In other words, knowing that J_2's sentence will scale with π, J_1 will convict unless the sentencer J_2 is at least twice as harsh. With such a sentencer, J_1 will acquit in all cases where the defendant has positive probability of guilt, instead of convicting in all those cases, exhibiting a jury nullification-like behavior to avoid the harsh punishment.

2.3 Model Extension 1: Mandatory Minimum

The baseline model assumes that there is no punishment inherent to the conviction itself. Something like a stigma attaching to merely being convicted would essentially create a minimum punishment $\underline{s} > 0$, unless J_i could also compensate defendants. This model extension considers the effect of a minimum punishment, which could be created by inherent costs of conviction or by legally mandated minimum sentences.

Since

$$EU_{v=1,s} - EU_{v=0} = 2\pi s s_i - s^2$$

J_i will prefer conviction if $s \in (0, 2\pi s_i)$.

Then, in the single-actor case, instead of always convicting when $\pi > 0$, J_i will

- convict and sentence to πs_i as in the baseline model if $\underline{s} \leq \pi s_i$ (the minimum is not binding)
- convict and sentence to $\underline{s}$ if $\pi s_i < \underline{s} < 2\pi s_i$
- acquit if $2\pi s_i \leq \underline{s}$ (assuming that J_i acquits when indifferent)

Rearranging the bounds on the above cases, a minimum $\underline{s}$ will bind when $\pi < \frac{\underline{s}}{s_i}$ and acquittal is preferred when $\pi \leq \frac{\underline{s}}{2s_i}$. Here, when the probability of guilt is low enough to bind, there is a range of π near 0 leading to acquittal (instead of conviction in the baseline model) followed by a range of π with conviction and increased sentences compared to the baseline model. Note that the jury nullification-like result in this single-actor extension depends on the probability of guilt — when the minimum binds, the expected sentence no longer can scale with π.

In the two-actor case, instead of convicting whenever $s_1 > \frac{s_2}{2}$, J_1 will

- convict knowing the sentence will be πs_2 when $s_1 > \frac{s_2}{2}$ and $\underline{s} < \pi s_2$ (the minimum is not binding);
- acquit knowing the sentence would be πs_2 when $s_1 \leq \frac{s_2}{2}$ and $\underline{s} < \pi s_2$ (the minimum is not binding);
- convict knowing the sentence will be $\underline{s}$ when $2\pi s_1 > \underline{s}$ and $\underline{s} \geq \pi s_2$;
- acquit knowing the sentence would be $\underline{s}$ when $2\pi s_1 \leq \underline{s}$ and $\underline{s} \geq \pi s_2$.

So, J_1 will behave identically to the baseline model when the minimum does not bind J_2. However, as in the single-actor case of this extension, J_1 sometimes will convict with a higher sentence (when $\frac{\underline{s}}{2s_1} < \pi < \frac{\underline{s}}{s_2}$). Note that these conditions imply that $s_1 > \frac{s_2}{2}$, and so J_1 would have convicted in the baseline model (with the lower sentence of πs_2) as well. Also as in the single-actor case, J_1 will acquit in cases when J_1 would have convicted in the baseline model. With a mandatory

minimum, J_1 will acquit when $\pi \leq \frac{\underline{s}}{2s_1}$ and also $\pi \leq \frac{\underline{s}}{s_2}$. Here, even if $s_1 > \frac{s_2}{2}$, leading to conviction in the baseline model, J_1 will acquit as long as $\pi \leq \frac{\underline{s}}{2s_1}$. Again, these additional acquittals stem from the minimum preventing the sentence from scaling with π.

2.4 Model Extension 2: Adding Lawfulness Utility

Now consider actors who care about abiding by the law. In this extension, an actor J_i loses utility when J_i deviates from either the legal probability threshold for guilt p_L or the sentence for the guilty s_L prescribed by the law. The sequences for the single- and two-actor version of this model extension differ from the baseline sequences only in that all actors know p_L and s_L at the beginning of the sequence.

This **lawfulness utility** stems from an actor who only cares about her own violations of each legal prescription separately — rather than, for example, about a social welfare function from which p_L and s_L are derived. This modeling choice might reflect actors who view their own rule-breaking in and of itself as a moral transgression or who fear repercussions for law-breaking.[10] Note that J_i suffers a lawfulness utility loss only when J_i is directly responsible for a law-violating step.

In the verdict step, when the defendant has probability of guilt π, J_i suffers the loss

$$-\alpha_i(\pi - p_L)^2$$

if J_i acquits even though $\pi > p_L$ or convicts even though $\pi \leq p_L$. In other words, an actor giving verdict v loses utility if applying the legal guilt threshold p_L would lead to a different verdict from v.

In the sentencing step, J_i suffers the loss

$$-\alpha_i(s - s_L)^2$$

when giving the sentence s.

J_i adds this lawfulness utility to the outcome utility in the baseline model. Both lawfulness utility losses are scaled by $\alpha_i \geq 0$, allowing for J_i to care more or less about following the law relative to the outcome of the case.

[10] Consider People v. Kriho (1999), reversing and remanding a lower court that found Kriho in contempt after other jurors reported that she had looked up the sentence for the charge online, stated that drug charges should be determined by "family and community," and handed out a pamphlet stating that jurors can vote according to conscience.

Single Actor: Sentencing (Step 2)

Because J_i controls both steps, J_i loses lawfulness utility for any deviations from either p_L or s_L, meaning that the expected utility from convicting and giving sentence s is

$$EU_{v=1,s} = (1-\pi)(-s^2) + (\pi)(-(s-s_i)^2) - \alpha_i(\pi - p_L)^2 \mathbb{1}_{\pi \leq p_L} - \alpha_i(s-s_L)^2$$

So, J_i will choose s^* to maximize this expected utility:

$$\begin{aligned} 0 &= (1-\pi)(-2s^*) + (\pi)(-2(s^* - s_i)) - 2\alpha_i(s^* - s_L) \\ s^* &= \frac{\pi s_i + \alpha_i s_L}{1+\alpha_i} \end{aligned}$$

In other words, J_i will choose a weighted average of the ideal sentence in the baseline model (πs_i) and the legal sentence (s_L). As J_i's concern for lawfulness decreases and α_i goes to 0, s^* goes to πs_i, and as α_i increases, s^* goes to s_L. As in the baseline case, unless $\pi s_i = s_L$, any conviction will lead to a sentence away from s_L.

Single Actor: Verdict (Step 1)

First, consider a simplified version of the model where the only lawfulness utility loss stems from the verdict. Then, the second step sentence chosen would be πs_i as in the baseline model. If the defendant's probability of guilt is under the legal threshold ($\pi \leq p_L$),

$$EU_{v=1,s=\pi s_i} - EU_{v=0} = \pi^2 s_i^2 - \alpha_i(\pi - p_L)^2$$

and instead of always convicting, J_i only convicts if

$$s_i^2 > \alpha_i \frac{(\pi - p_L)^2}{\pi^2}$$

In contrast, if the defendant's probability of guilt is over the legal threshold ($\pi > p_L$),

$$EU_{v=1,s=\pi s_i} - EU_{v=0} = \pi^2 s_i^2 + \alpha_i(\pi - p_L)^2$$

and J_i will always convict, as in the baseline model. Here, J_i and the law both favor conviction, so J_i's behavior is unaffected. In other words, the net expected utility for J_i of convicting rather than acquitting is translated up when the law prefers conviction and down when the law prefers acquittal, as shown in Figure 2.1.

Second, consider a different simplified version in which the only lawfulness utility loss stems from the sentence. Then, the chosen sentence will be the weighted

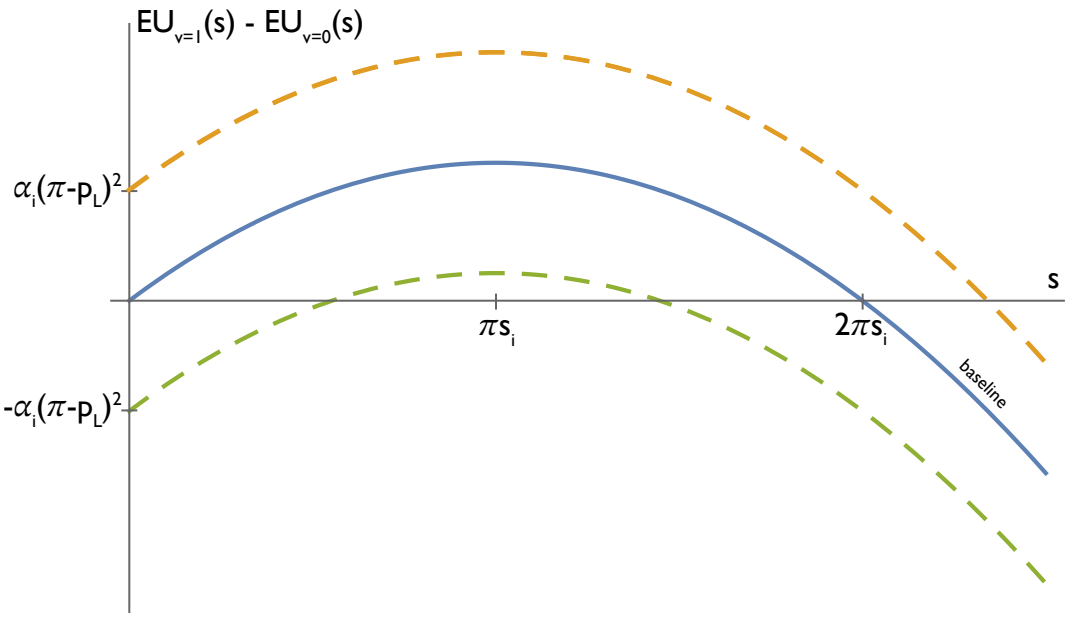


Figure 2.1: An example of net expected utility of convicting rather than acquitting in the baseline model and with lawfulness utility loss only from the verdict, both where the law favors conviction and where it favors acquittal.

average of πs_i and s_L noted above $\left(s^* = \frac{\pi s_i + \alpha_i s_L}{1+\alpha_i}\right)$, and

$$\begin{aligned} EU_{v=1,s=s^*} - EU_{v=0} &= 2\pi s^* s_i - (s^*)^2 - \alpha_i (s^* - s_L)^2 \\ &= \frac{\pi^2 s_i^2 + 2\alpha_i \pi s_i s_L - \alpha_i s_L^2}{1+\alpha_i} \end{aligned}$$

and J_i will convict if

$$\begin{aligned} \pi^2 s_i^2 + 2\alpha_i \pi s_i s_L - \alpha_i s_L^2 &> 0 \\ \pi s_i &> (\sqrt{\alpha_i^2 + \alpha_i} - \alpha_i) s_L \end{aligned}$$

The right-hand side is 0 for $\alpha_i = 0$ and approaches $\frac{s_L}{2}$ as $\alpha_i \to \infty$. So, as in the baseline model, J_i always convicts when the weight on lawfulness is zero, and approaches convicting only if s_L is less than $2\pi s_i$, which as in the baseline model is a threshold between preferring to convict and preferring to acquit. As shown in Figure 2.2, lawfulness utility loss from sentencing shifts the optimal sentence from πs_i towards s_L; because the loss is incurred only when the defendant is convicted, it translates the net expected utility of convicting down; and because the loss increases away from s_L, it also narrows the range of sentences for which conviction is preferred.

Combining both types of lawfulness utility losses simply means adding or subtracting the loss from the verdict $\alpha_i(\pi - p_L)^2$ to the net expected utility of convicting

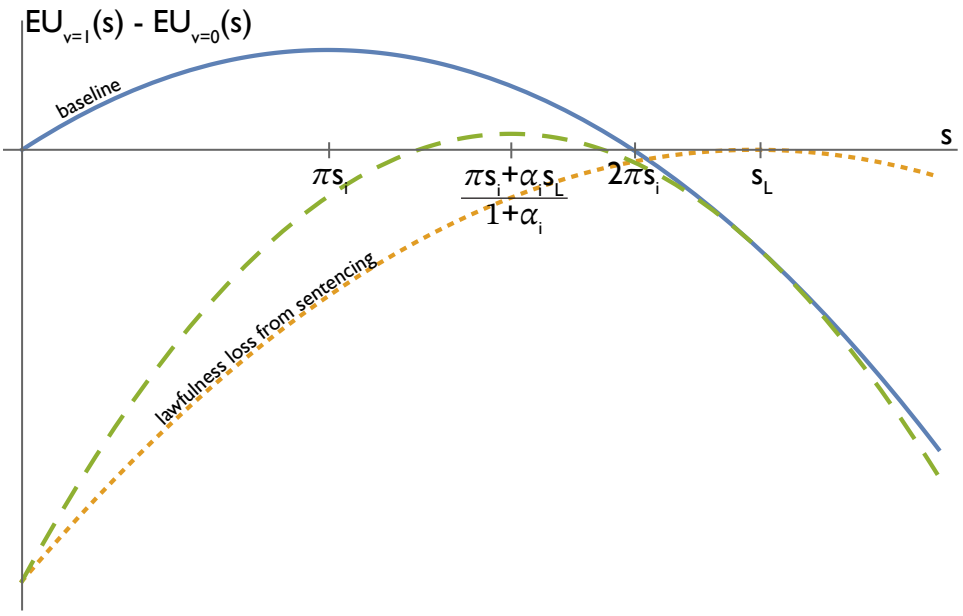


Figure 2.2: An example of net expected utility of convicting rather than acquitting in the baseline model, lawfulness utility loss from sentencing, and their sum.

when there is only loss from the sentencing:

$$EU_{v=1,s=s^*} - EU_{v=0} = \frac{\pi^2 s_i^2 + 2\alpha_i \pi s_i s_L - \alpha_i s_L^2}{1+\alpha_i} \pm \alpha_i(\pi - p_L)^2$$

where the last term is added when the law favors conviction and subtracted when it favors acquittal. Then, J_i convicts when $\pi > p_L$ if

$$s_L < \pi - p_L$$

or if $s_L \geq \pi - p_L$ and

$$\pi s_i > \sqrt{\alpha_i^2 + \alpha_i}\sqrt{s_L^2 - (\pi - p_L)^2} - \alpha_i s_L$$

In addition, J_i convicts when $\pi \leq p_L$ if

$$\pi s_i > \sqrt{\alpha_i^2 + \alpha_i}\sqrt{s_L^2 + (\pi - p_L)^2} - \alpha_i s_L$$

Graphically, the combined lawfulness utility means taking the net expected utility with only losses from sentencing (the green dashed line in Figure 2.2), and translating it up if $\pi > p_L$ and down if $\pi \leq p_L$.

If the law favors acquittal and the lawfulness cost from convicting instead is larger than $\pi^2 s_i^2$, then J_i will always acquit—the baseline curve will lie below zero, and

further losses due to sentencing would only bring it further down. Similarly, if s_L is large enough, J_i also will always acquit, even if the law favors conviction, since $\sqrt{\alpha_i^2 + \alpha_i}\sqrt{s_L^2 - (\pi - p_L)^2} - \alpha_i s_L$ will approach $(\sqrt{\alpha_i^2 + \alpha_i} - \alpha_i)s_L$. This nullification-like behavior stems from the lawfulness cost of sentencing away from s_L outweighing the lawfulness cost of acquitting instead of convicting. So, these lawfulness costs can push an actor to acquit instead of always convicting as in the baseline case, because of the cost of convicting itself when $\pi \leq p_L$, because of the cost of sentencing away from s_L after a conviction, or because of a combination of these two costs.

Two Actors: Sentencing (Step 2)

The analysis here is the same as for the single-actor case. Although the $-\alpha_i(\pi - p_L)^2 \mathbb{1}_{\pi \leq p_L}$ term is no longer in the expected utility (because J_2 is not responsible for determining the verdict), that term did not affect the maximization over s. So, s^* will be chosen as before, with $i = 2$:

$$s^* = \frac{\pi s_2 + \alpha_2 s_L}{1 + \alpha_2}$$

Note that J_2 will give this sentence even if J_2 would not have chosen to convict the defendant; J_2 would not, for example, give a sentence of 0 when J_2 would have acquitted in a single-actor process. Giving a lower sentence would increase the lawfulness cost from the sentence. If lawfulness costs for sentencing took into account the preferred verdict under the law—for example, by having $s_L = 0$ when $\pi \leq p_L$—then such reduced sentences could occur. However, the legal remedy for an incorrect conviction is not a reduced sentence, and the separation of these steps in the law is reflected in the independent verdict and sentencing costs in the model.

Two Actors: Verdict (Step 1)

Unlike in the single-actor case, J_1 pays a lawfulness cost only for an unlawful verdict, and not for any deviation in the sentence upon conviction. So, if the defendant's probability of guilt is under the legal threshold ($\pi \leq p_L$),

$$EU_{v=1,s=s^*} - EU_{v=0} = 2\pi s^* s_1 - (s^*)^2 - \alpha_1(\pi - p_L)^2$$

and if the defendant's probability of guilt is over the legal threshold ($\pi > p_L$),

$$EU_{v=1,s=s^*} - EU_{v=0} = 2\pi s^* s_1 - (s^*)^2 + \alpha_1(\pi - p_L)^2$$

Note that there is a difference between the single-actor case and a two-actor process with two identical actors. Because the single actor J_i is responsible for both steps in the process, J_i pays lawfulness costs related to both the verdict and the sentence. While this difference does not affect the sentence chosen upon conviction, it means that $(EU_{v=1,s=s^*} - EU_{v=0})$ for the single actor J_i is equal to those above for first-step actor J_1 minus the sentencing cost $\alpha_1(s^* - s_L)^2$. As a result, as long as $s^* \neq s_L$, the process with two identical actors will convict over a broader range of the parameter space than one of those actors would if responsible for both verdict and sentence. The first-step, verdict-choosing J_1 can pass off the cost of sentencing away from the lawful sentence to J_2. This kind of free-riding effect will occur as long as actors in the process have preferences specific to their own actions.

J_1 will convict if $\pi \leq p_L$ when

$$2\pi s_1 > s^* + \frac{\alpha_1(\pi - p_L)^2}{s^*}$$
$$s_1 > \frac{(\pi s_2 + \alpha_2 s_L)^2 + \alpha_1(1+\alpha_2)^2(\pi - p_L)^2}{2\pi(1+\alpha_2)(\pi s_2 + \alpha_2 s_L)}$$

and will convict if $\pi > p_L$ when

$$2\pi s_1 > s^* - \frac{\alpha_1(\pi - p_L)^2}{s^*}$$
$$s_1 > \frac{(\pi s_2 + \alpha_2 s_L)^2 - \alpha_1(1+\alpha_2)^2(\pi - p_L)^2}{2\pi(1+\alpha_2)(\pi s_2 + \alpha_2 s_L)}$$

2.5 Comparative Statics

Baseline Model

In the single-actor baseline model with only outcome utility, J_i always convicts and gives the sentence πs_i. So, changing parameters has no effect on the verdict, but increasing the probability of guilt π or J_i's preferred sentence for the guilty s_i strictly increases the sentence given.

In the two-actor baseline model, J_1 convicts if $2s_1 > s_2$, and J_2 gives the sentence πs_2. As in the single-actor case, increasing π or s_2 strictly increases the sentence given upon conviction. However, increasing s_2 across the threshold of $2s_1$ changes the verdict from conviction to acquittal. So, raising s_2 from 0 will first lead to convictions with increasing sentences, but at the threshold of $2s_1$ will switch to acquittals as J_1 reacts to J_2's relative harshness. The probability of guilt π does not affect the verdict since both J_1 and J_2 scale their preferred sentence with π.

Model Extension 1: Mandatory Minimum

Model Extension 1: Single-Actor

In the single-actor extension with a mandatory minimum, increasing π or s_i weakly increase the sentence and can move the verdict from acquittal to conviction. Increasing the minimum $\underline{s}$ also weakly increases the sentence, but shifts the verdict from conviction to acquittal across the threshold $2\pi s_i$.

Model Extension 1: Two-Actor

In the two-actor extension with a mandatory minimum, increasing π, s_2, or $\underline{s}$ weakly increases the sentence.

Increasing s_1 can shift the verdict from acquittal to conviction, but will not affect the sentence given.

The effect of increasing s_2 on the verdict depends on whether $2\pi s_1$ is greater than $\underline{s}$. First, if it is, then for low values of s_2, J_1 will convict with a sentence of $\underline{s}$. As s_2 increases so that πs_2 becomes slightly greater than $\underline{s}$, J_1 will convict with a sentence of πs_2 since $2\pi s_1 > \underline{s} \approx \pi s_2$. However, as s_2 continues to increase, J_1 will switch to acquitting to avoid the high πs_2 sentence. Second, if $2\pi s_1 \leq \underline{s}$, then J_1 acquits for low s_2. When s_2 increases so that $\underline{s} < \pi s_2$, it follows that $2\pi s_1 \leq \underline{s} < \pi s_2$ and so $s_1 < \frac{s_2}{2}$. So, J_1 continues to acquit.

The effect of increasing $\underline{s}$ on the verdict depends on whether $2\pi s_1 \leq \pi s_2$. First, if so, for low $\underline{s}$, J_1 will acquit instead of convicting with πs_2. As $\underline{s}$ increases to be greater than πs_2, since $2\pi s_1 \leq \pi s_2$, $2\pi s_1$ will also be less than $\underline{s}$. J_1 will acquit instead of convicting with $\underline{s}$. Second, if $2\pi s_1 > \pi s_2$, then for low $\underline{s}$, J_1 will convict with πs_2. As $\underline{s}$ increases to be slightly greater than πs_2, J_1 will convict with $\underline{s}$ since $2\pi s_1 > \pi s_2 \approx \underline{s}$. However, as $\underline{s}$ continues to increase, J_1 will switch to acquitting instead of convicting with $\underline{s}$.

To see the effect of increasing π on the verdict, consider two cases. First, if $2\pi s_1 \leq \pi s_2$, when π is small J_1 will acquit instead of convicting with a sentence of $\underline{s}$. As π increases, πs_2 will become greater than $\underline{s}$ before $2\pi s_1$. In other words, the mandatory minimum will stop binding before the condition for convicting with $\underline{s}$ will be met. So, J_1 will continue to acquit, now instead of convicting with a sentence of πs_2. Second, if $2\pi s_1 > \pi s_2$, J_1 still acquits instead of convicting with $\underline{s}$ for small π. As π increases so that $2\pi s_1 > \underline{s}$, J_1 will convict with $\underline{s}$, and finally, as π continues to increase so that $\pi s_2 > \underline{s}$. J_1 will convict with πs_2.

Model Extension 2: Adding Lawfulness Utility

Model Extension 2: Single-Actor

In the single-actor extension with lawfulness, increasing π, s_i, or s_L increases the sentence, while increasing α_i shifts the sentence from πs_i to s_L.

Increasing π increases the net expected utility of conviction. Note that the verdict lawfulness cost is subtracted from the net expected utility when $\pi \leq p_L$, and so increasing π reduces the cost. When $\pi > p_L$ increasing π increases the lawfulness cost, but the law prefers conviction.

Increasing p_L decreases the net expected utility of conviction. For low p_L, the law favors conviction, and the lawfulness benefit of convicting decreases as p_L increases to approach π. As p_L increases further, the law favors acquittal, and the lawfulness cost of convicting increases.

Increasing s_i increases the net expected utility of conviction and will eventually make the net expected utility positive.

The partial derivative of the net expected utility with respect to s_L is

$$\frac{2\alpha_i \pi s_i - 2\alpha_i s_L}{1 + \alpha_i}$$

Increasing s_L when $s_L \leq \pi s_i$ also increases the net expected utility of conviction, but continuing to increase s_L then decreases the net expected utility. In other words, the net expected utility of conviction increases as the lawful sentence s_L gets closer to J_i's preferred sentence πs_i.

The partial derivative of net expected utility with respect to α_i is

$$\frac{-(\pi s_i - s_L)^2}{(1 + \alpha_i)^2} \pm (\pi - p_L)^2$$

where the last term is added when $\pi > p_L$ and subtracted when $\pi \leq p_L$. So, when the law favors acquittal ($\pi \leq p_L$), increasing α_i always reduces the net expected utility of conviction because increased weight on lawfulness increases the cost of convicting and also makes the sentence upon conviction further from πs_i. When the law favors conviction ($\pi > p_L$), increasing α_i increases the net expected utility if $1 + \alpha_i > \frac{|\pi s_i - s_L|}{\pi - p_L}$. So, if $|\pi s_i - s_L| < \pi - p_L$, increasing α_i always increases net expected utility. If $|\pi s_i - s_L| > \pi - p_L$, then increasing α_i from 0 will decrease net expected utility until $\alpha_i = \frac{|\pi s_i - s_L|}{\pi - p_L} - 1$ and then will increase net expected utility.

Model Extension 2: Two-Actor

In the two-actor extension with lawfulness, increasing π, s_2, or s_L increases the sentence, while increasing α_2 shifts the sentence from πs_2 to s_L.

Because α_1 only enters into J_1's net expected utility of conviction through the lawfulness cost of the verdict, as α_1 increases, net expected utility of conviction decreases when the law favors acquittal ($\pi \leq p_L$) and increases when the law favors conviction ($\pi > p_L$).

Note that α_2 enters into the net expected utility only through s^*, and as α_2 increases from 0, s^* moves from πs_2 to s_L, increasing or decreasing monotonically. Since J_1's net expected utility is $2\pi s^* s_1 - (s^*)^2 \pm \alpha_1(\pi - p_L)^2$, it is maximized when $s^* = \pi s_1$. Thus, the effect of increasing α_2 on net expected utility depends on the relative values of πs_1, πs_2, and s_L. If s_L is in between πs_1 and πs_2, then increasing α_2 will increase net expected utility. If πs_2 is in the middle, then increasing α_2 will decrease net expected utility. If πs_1 is in the middle, then increasing α_2 until $s^* = \pi s_1$ will increase net expected utility, but further increases will decrease net expected utility.

The partial derivative of the net expected utility with respect to p_L is

$$2\alpha_1(\pi - p_L)$$

when $\pi \leq p_L$, and

$$-2\alpha_1(\pi - p_L)$$

when $\pi > p_L$. Increasing p_L always decreases net expected utility.

Because s_1 only enters into the first term of the net expected utility and not into s^*, increasing s_1 increases net expected utility.

The partial derivative of the net expected utility with respect to s_2 is

$$\frac{2\pi(\pi s_1 - \pi s_2) + 2\alpha_2\pi(\pi s_1 - s_L)}{(1+\alpha_2)^2}$$

So, increasing s_2 increases net expected utility if

$$(\pi s_1 - \pi s_2) + \alpha_2(\pi s_1 - s_L) > 0$$

Then, if $\pi s_1 < \frac{\alpha_2}{1+\alpha_2} s_L$, increasing s_2 will always decrease net expected utility. For higher values of πs_1, increasing s_2 will begin by increasing net expected utility, but once $\pi s_2 \geq \pi s_1 + \alpha_2(\pi s_1 - s_L)$ further increases will decrease net expected utility.

The partial derivative of net expected utility with respect to s_L is

$$\frac{2\alpha_2(\pi s_1 - \pi s_2) + 2\alpha_2^2(\pi s_1 - s_L)}{(1+\alpha_2)^2}$$

So, as for s_2, increasing s_L increases net expected utility if

$$(\pi s_1 - \pi s_2) + \alpha_2(\pi s_1 - s_L) > 0$$

If $s_1 < \frac{1}{1+\alpha_2} s_2$, increasing s_L will always decrease net expected utility. For higher values of s_1, increasing s_L will begin by increasing net expected utility, but once $s_L > \frac{1+\alpha_2}{\alpha_2}\pi s_1 - \frac{1}{\alpha_2} s_2$ further increases decrease net expected utility.

Non-Monotonic Comparative Statics in the Lawfulness Utility Extension

Increasing Lawfulness in a Single-Actor Process

Consider the effect of increasing the lawfulness of a single actor deciding verdict and sentence. Note that when $\alpha_i = 0$ the actor behaves as in the baseline model, convicting and giving a sentence of πs_i. As $\alpha_i \to \infty$, the actor's verdict and, if relevant, sentence will approach the lawful ones. Thus, an infinitely lawful actor will convict and give s_L when the law favors conviction ($\pi > p_L$). So, when the law favors conviction, extreme values of α_i will always lead to conviction.

As noted above, if $|\pi s_i - s_L| > \pi - p_L$, then increasing α_i from 0 will decrease the net expected utility of conviction until $\alpha_i = \frac{|\pi s_i - s_L|}{\pi - p_L} - 1$ and then will increase net expected utility. In other words, if the sentencing lawfulness cost is relatively large compared to the verdict lawfulness cost, then the benefit of conviction reaches a minimum for some intermediate level of lawfulness. A scenario where, for example, $s_L >> \pi s_i$ can make this reduction in net expected utility large enough to flip the actor's verdict from conviction to acquittal. A completely lawless actor would convict and give a small punishment, a moderately lawful one would nullify and acquit, and a completely lawful one would convict and give the law's harsh penalty.

Increasing the Expected Sentence Due to Factors Outside of the First Actor's Control

Several parameters have a non-monotonic effect on the net expected utility that goes in the opposite direction, increasing and then decreasing the benefit of conviction so that it is maximized for an intermediate value of the parameter. This effect occurs when increasing the parameter causes the expected sentence to increase over an interval that includes the baseline sentence of the actor deciding the verdict (πs_i in a single-actor process or πs_1 in a two-actor one).

In a single-actor process, this effect occurs when increasing s_L, when the law also prefers acquittal ($\pi \leq p_L$). When the lawful sentence is zero, the lawfulness cost of convicting may outweigh the benefit from the optimal sentence. As the lawful sentence increases to the actor's preferred sentence, the chosen sentence will also approach the actor's preferred sentence, and the benefits can increase to make conviction desirable. However, as the lawful sentence continues to increase, the actor eventually will prefer a nullification-like acquittal.

In the two-actor process, this pattern in the net expected utility when increasing s_L only occurs if $\pi s_1 > \frac{\pi}{1+\alpha_2} s_2$. The right-hand side $\frac{\pi}{1+\alpha_2} s_2$ is the sentencer's chosen sentence s^* when $s_L = 0$. Since increasing s_L will also increase s^* (assuming a non-zero lawfulness α_2), this condition ensures that s^* will start below πs_1 when $s_L = 0$ and so will eventually increase past πs_1 as s_L increases. Note that in the single-actor case, s^* will always pass through πs_i as s_L increases.

The two-actor process also sees this pattern when increasing the lawfulness of *the sentencer* α_2, when πs_1 sits between πs_2 and s_L. Because raising α_2 will move s^* from πs_2 to s_L, the same pattern of acquittal, conviction, and then acquittal again can occur.

2.6 Empirical Implications

Impact of Judge Harshness on Conviction Rates

Related empirical literature discussed above has examined the nullification and compromise verdict outcomes predicted by the model. However, the focus has been on harsh sentences set across a jurisdiction through sentencing guidelines or mandatory minimums. This model's choice to consider the sentencer as a separate actor with a separate utility function suggests an additional avenue for empirical exploration: does the harshness of a particular judge reduce conviction rates?

Kaplan & Krupa (1986) provides experimental evidence with student subjects who believed they were voting on real punishments, finding the lowest conviction rate occurred when students were told an authority would choose a punishment in the moderate-to-severe range. Students also gave more severe punishments when they controlled the punishment than they recommended to an authority in charge of the sentence. In this setting, it seems likely that students would assume they were more lenient than the authority. The variation in public opinion and judge sentencing choices suggests that real-life juries sometimes are more harsh than the judges in their cases, in addition to sometimes being more lenient.

A large literature beginning with Kling (2006) has taken advantage of the random assignment of cases to judges to use judge harshness — measured by the average sentence given by a judge — as an instrumental variable to see how the sentence given to a particular offender affects other outcomes. Here, the random assignment may allow for analysis of how these harshness levels affect jury nullification decisions. One requirement for an effect to be seen is that juries can perceive the harshness of judges during the trial process. Blanck (1993) states "[i]n a criminal trial, a trial judge's beliefs or expectations for a defendant's guilt may be manifested either verbally or nonverbally . . . and can be reflected in a judge's comments on evidence, responses to witness testimony, reactions to counsels' actions, or in rulings on objections."[11] Dietrich et al. (2019) examine oral argument audio recordings and find that "*vocal pitch alone* is strongly predictive of Supreme Court Justices' votes" and "nonsubstantive and implicit signals, even among elite actors such as federal judges and Supreme Court Justices, can provide additional meaningful information on their attitudes beyond what can be found in their textual pronouncements." These studies suggest that information about judges' relative harshness may leak to jury even though the jury is not given information about the expected sentence in the case. In addition, if it were possible also to assess jury harshness and find cases with both lenient juries and judges, the free-riding effect could be supported empirically.

Impact of Changing Lawfulness on Conviction Rates

Jury instructions may be able to toggle the lawfulness of juries deciding sentences. Grover (2019) argues that capital juries are not treated or instructed in a manner that engenders jurors to feel responsible for the sentences they give. Horowitz (2008) describes an earlier experiment with mock jurors examining the effect of instructions explicitly telling them they have a right to nullify. Although juries "did not often explicitly admit that they were ignoring the law . . . they tended to construe the evidence differently so as to support their verdicts," and this effect was weakened if jurors also were reminded that they should follow the law. Horowitz notes a later explanation for this behavior in Diamond (2007): the nullification instruction "implicitly released the jurors from the yoke of legal obligation that ordinarily ties their decisions closely to the legal requirements outlined in the other jury instructions."

[11] Observations of judge behavior were used to categorize judges on four dimensions: judicial ("professional, wise, competent, and honest"), directive ("dogmatic and dominant"), confident ("less anxious and less hostile"), and warm ("open-minded and empathic"), and found that convictions were correlated with judges who were less judicial, less directive, and more engaged.

Judges' lawfulness may be affected by the conditions for their reappointment or reelection. For example, a Virginia legislator questioned judges up for reappointment about deviations above the suggested sentencing guidelines.[12] (Hurston, 2023) Cohen et al. (2015) finds that — in the period before elections — lenient judges give harsher sentences and harsh judges give more lenient ones, suggesting that judges move away from their own personal preferences when nearing re-election.

Thus, changes in jury instructions in jury sentencing jurisdictions and changes in judges reputation-based incentives may help investigate the effects of lawfulness on conviction rates. However, for both juries and judges, it may be difficult to get enough variation in lawfulness to observe the full non-monotonic conviction rate behavior predicted by the model.

2.7 Conclusion

This paper creates a game theoretic model of the determination of a verdict and sentence. It relates to a literature of models considering the effect of changes in sentencing on convictions. However, unlike most prior models, it includes a trial-level actor — a judge or a jury — as the originator of the sentence, rather than a higher-level authority setting broad policy. In doing so, the model allows for novel analysis related to that actor's lawfulness, a quality often referenced in the legal scholarship debating the merits of juries versus judges as decisionmakers. By focusing entirely on trial-level actors, the model also enables comparisons between the typical two-actor (jury then judge) process and single-actor processes (bench trials

12

> Surovell, a noted advocate for criminal justice reform, asked two judges about their record of sentencing convicts above the suggested guidelines.
>
> Since 2022, Judge S. Anderson Nelson of the 10th Circuit Court has doubled his rate of departure from the guidelines, the senator noted.
>
> "I've had some pretty bad cases, murder and some horrendous rape cases," Nelson said. "Juries were going above the guidelines, like the last trial where the jury tripled the guidelines sentence."
>
> On the other hand, Surovell said Judge Ricardo Rigual of the 15th Circuit Court had an unusual pattern of sentencing.
>
> "When you were first on the bench, you were above the guidelines 27% of the time and 6% below," he said. "Last year, you were above 15% and below 2%,"
>
> Rigual asked the legislators to look at his sentencing for bench trials where he's 100% in charge.
>
> "I'm right there with the rest of the state," the judge pointed out. "Spotsylvania County has a great deal of plea agreements that often go above the guidelines if they drop other charges."

(Hurston, 2023)

and jury sentencing). These modeling choices generate new empirical predictions about the effect of these institutional design choices on the conviction rate.

The single-actor baseline version of the model, in which the actor only cares about the outcome of the case and not lawfulness, yield a stark result: convictions in all cases, accompanied by sentences varying with the probability of guilt. With full sentencing discretion, nothing prevents the actor from providing incremental sentences for defendants who are unlikely to be guilty. The starkness of this result highlights that acquittals can stem from 1) restrictions on sentencing discretion — a mandatory minimum or another actor controlling the sentence — or 2) preferences that are not outcome-based. Acquittals from restrictions on sentencing discretion are example of jury nullification, and these acquittals only correlate with the probability of guilt if the expected sentence does not, either because it is fixed or because the sentencer has preferences that are not outcome-based.

When actors also care about their own lawfulness, several additional effects occur. First, if the jury and judge in a two-actor process both agree that the lawful sentence is too harsh, a free-riding effect can increase convictions in the two-actor process as compared to the single-actor version. Second, increasing lawfulness and increasing the sentence required by law can cause the actor responsible for the verdict to switch from one verdict to the other, but further increases can cause a switch back. Actors who are completely lawless and those who are completely lawful may make the same decision, while moderately lawful actors do not.

As is intuitive, if the actors' lawfulness can be increased to an extreme, the lawful verdict and sentence will be achieved. However, policies like alterations to jury instructions and pressures on judges from reappointment or election have not driven these actors to be completely lawful, and they seem unlikely to be able to do so. These policies instead shift lawfulness within the intermediate range, and the model's results suggest that such shifts can cause the verdict to change *away* from the lawful one.

In future work, the model's two-actor structure would allow for evaluation of policies that give (or hide) information about the expected sentence to the jury. U.S. jurisdictions generally have attempted to prevent 1) the verdict from being influenced by the expected sentence and 2) the sentence from being affected by the probability of guilt. Courts almost uniformly attempt to prevent the jury from learning about potential sentences for noncapital crimes — an "iron law of jury ignorance" reinforced through jury instructions. (Epps & Ortman, 2022) In addition, noncapital sentencing

factors do not include whether the defendant was more likely to have committed the crime. The aim is "a binary all-or-nothing criminal sentencing regime — where no punishment is imposed . . . below the 'reasonable doubt' threshold, while from that point and on, punishment . . . is disconnected from the probability of guilt" (Fisher, 2011).

Many legal scholars argue that juries ought to be better informed about the sentencing consequences of their verdicts. For example, Bellin (2010) advocates for attorneys to argue that information about harsh sentences should be admissible as relevant evidence concerning the probability of guilt because a "defendant's ex ante awareness of a severe punishment for a charged crime — corroborated, or in some cases established, by a simple description of the applicable sentencing law — decreases somewhat the probability that he committed that offense." In another example, Epps & Ortman (2022) propose that jury instructions should include the minimum and maximum sentences under the law as well as whether sentences would be consecutive, arguing that informing juries can only push towards leniency. However, Virginia juries involved in jury sentencing were given similar instructions to this proposal and were significantly harsher than judges. In one example, a defendant faced three counts, each with a five-year minimum and forty-year maximum sentence. The jury was informed of these ranges, but uninformed about Virginia sentencing guidelines, which would have recommended a total sentence between six years and four months and ten years and five months. The jury instead recommended a sixty-five year sentence. (McCloy, 2021; Klein, 2021) It appears possible that they anchored around the average of the minimum-maximum range, giving around twenty years for each count. Although it is uncertain what this jury would have chosen under other informational conditions, it seems possible that the information given to them can be highly influential.

The model discussed in this paper assumes that the first actor has full information about a second actor's sentence and about the lawful sentence. An extension of the model to allow for uncertainty about these sentences would allow for comparisons between different informational environments. Empirical testing might be possible through consideration of changes in jury instructions or of certain types of cases that involve high-profile mandatory sentences, such that jurors are more likely to be aware of the expected sentence upon conviction.

Chapter 3

IMPLEMENTING INFORMATION ESCROWS

with Alexander V. Hirsch

Information escrows keep reports of information secret until certain conditions are met. As Ayres & Unkovic (2012) describe, such escrows may increase the amount of information transmitted. For example, an information escrow that holds reports in escrow until more than one of the same kind of report is made can motivate people to make reports when they otherwise would have been deterred by a "first-mover disadvantage to unintermediated communication."

The canonical example of an information escrow is one that escrows allegations of misconduct until there are multiple reports filed against the same person. People who experience, for example, sexual harassment from a powerful person may be unwilling to report because of fears of retaliation, as well as fears that they will not be believed. In these cases, people may become willing to report if they know that their report only will become public when another — or several other — people accuse the same perpetrator. This kind of information escrow threshold may reduce the amount of possible retaliation or at least spread the cost among multiple accusers, and it can increase others' beliefs that the accusation was truthful.

Escrows for misconduct allegations are not purely theoretical. The most prominent implementation is Callisto,[1] a platform for private reporting of campus sexual assault. Callisto is a nonprofit that aims to "empower survivors of sexual assault, provide a safe alternative to reporting, and increase the likelihood that serial perpetrators will be held accountable." The Callisto website accepts reports of campus sexual assault and, if two or more reports name the same perpetrator, Callisto has an attorney notify the reporters of the existence of multiple reports and offer them legal options counseling. In January 2024, the American Economic Association launched a Reporting Lockbox with a similar design to Callisto: if two or more members log an incident involving the same person, they are contacted to see if they would be interested in coordinating with each other.[2] The American Political

[1] https://www.projectcallisto.org

[2] https://www.aeaweb.org/about-aea/reporting-lockbox; https://www.aeaweb.org/news/member-announcements/2024-jan-18

Science Association has considered similar measures (APSA Council, 2018).

The design of Callisto and the Reporting Lockbox involves many choices: for example, the numerical threshold for releasing information (two reports) and the ability to withdraw reports or reject coordinated action with other reporters. Yet, the theory behind escrows remains largely unexplored.

In this paper, we take a mechanism design approach to a particular setting for an information escrow: a firm seeking to minimize misconduct by a manager, perpetrated against the firm's other employees. A manager faces the immediate consequence of firing if the escrow's threshold is satisfied. In this setting, we find that the employees' cost for reporting misconduct is of central importance to the existence of — and the firm's utility from — a truthful mechanism where the firm commits to fire the manager if the number of misconduct reports exceeds a threshold. We also find that, in general, single-threshold mechanisms will be non-optimal when it is costly to report misconduct.

We consider two versions of the model: one in which all reports are costless, and one in which reporting misconduct is costly while reporting that no misconduct occurred is costless. In the second version, we focus on this asymmetric cost of reporting as a reflection of settings where there is an inherent cost in reporting misconduct, such as the cost of detailing and reliving a traumatic interaction.

An employee under an escrow system only affects the firing decision when *pivotal*. In other words, the employee's report only matters if the number of misconduct reports from all other employees is exactly at the threshold for the escrow. Then, one more misconduct report will lead the firm to fire the manager, while one more report of no misconduct will lead the firm to retain the manager. The employee then decides whether to report misconduct by considering the net benefit of reporting misconduct versus not, conditional on the employee being pivotal and weighted by the probability that the employee is pivotal. Thus, when reports are costless, this setting is parallel to the jury voting in Austen-Smith & Banks (1996), and just as aggregation under majority rule leads to truthful voting in the jury setting, an information escrow with a threshold set optimally for the firm supports truthful reporting of misconduct, as the threshold is also optimal for the employees. This single-threshold mechanism therefore will give the firm its first-best outcome.

In contrast, when reporting misconduct is costly, the firm's optimal threshold may no longer make truthful reporting incentive compatible. Now, truthful reporting

requires that the net benefit of reporting misconduct (still conditional on pivotality and weighted by the probability of being pivotal) be greater than the reporting cost when the employee experienced misconduct, but less than the reporting cost when the employee did not. Because pivotality depends on the threshold, these net benefit values both change with the threshold, and each threshold determines its own range of reporting costs for which truthful reporting will be incentive compatible. In addition, while an employee who experiences misconduct always believes the manager is worse than an employee who does not, the probability of being pivotal can be higher for an employee who does not experience misconduct. That difference in pivot probabilities can be great enough that the net benefit of falsely reporting misconduct exceeds the net benefit of truthfully reporting it for a given threshold, meaning that no truthful single-threshold mechanism exists for that threshold and any reporting cost. It is possible for the thresholds that permit truthful mechanisms (when paired with some reporting cost) to not be consecutive, and it is possible for the reporting costs that permit truthful mechanisms (when paired with some threshold) to not form a continuous interval.

Finally, single-threshold mechanisms are generally non-optimal in a setting with costly reporting. First, if the firm's prior belief is that the manager is worse than the pool, using the best possible single-threshold mechanism may still be worse than always firing the manager on the basis of the prior. The possible truthful single-threshold mechanisms may have thresholds so high that the firm prefers to always fire the manager rather than to commit to retaining the manager when the number of misconduct reports is below the threshold but above the firm's first-best cutoff. Second, the firm can improve its outcome by mixing over two thresholds, for example by committing to use one threshold with probability p and another with probability $(1 - p)$. Mixing allows the firm to set the net benefit of truthfully reporting misconduct to equal precisely the cost of reporting, rather than restricting the firm to the net benefits corresponding to integer thresholds. Furthermore, if the reporting costs permitting truthful reporting do not form a continuous interval, mixing can allow the firm to create a truthful mechanism for a reporting cost in the gap. With mixed thresholds, the firm can elicit true information when it could not under any single threshold.

3.1 Related Literature

This work adds to a growing recent literature considering the effects of information escrows on the reporting of harassment.

Lee & Suen (2020) center their analysis on the possible existence of libelers making false accusations. In their model, libelers only arise for innocent agents who never harass, while guilty agents have some probability of harassing and creating a victim. They find that libelers are more likely to delay reporting than victims because victims have a stronger belief that their reports will be corroborated later. Applying this framework to an information escrow, they note that an escrow may deter reporting because the escrow reduces the cost of immediate reporting for libelers, making reports through the escrow less credible.

This paper does not include actors, like the libelers, with a particular animus against the manager due to reasons other than the manager's tendency to commit misconduct. The potential for false reports of misconduct in our model stems from employees who did not experience misconduct having higher probabilities of being pivotal, and we focus on truthful mechanisms. The widespread belief that underreporting is common due to retaliation and reputation costs for reporters suggests that eliciting information from people who experienced misconduct is a significant problem on its own (Barak-Corren & Lewinsohn-Zamir, 2019; Tuerkheimer, 2019; Dobbin & Kalev, 2020).

Pei & Strulovici (2021) take a mechanism design approach and consider a strategic perpetrator as well as reporting agents who may have an animus against that perpetrator leading to false reports. Here, the main conclusion from witnessing an offense in equilibrium under an information escrow that requires two offenses for punishment is knowing that the perpetrator will be deterred from committing a second offense and triggering the escrow.

In our paper, we take managers not to be strategic, assuming that they will perpetrate misconduct with a certain probability regardless of the escrow threshold or other system design choices. Serial harassment and assault appear common. Widman et al. (2013) survey a small sample of convicted sex offenders and demographically comparable community men and find that 53% of sex offenders and 47% of community men admit to more than one act of sexual assault since turning 14 years old, and 25% of sex offenders and 28% of community men admit to five or more acts. Cantalupo & Kidder (2018) note the prevalence of serial harassers among cases of sexual harassment by faculty. Lovell et al. (2018) and Hagerty (2019) report on the results of testing thousands of previously untested rape kits in Cuyahoga County, revealing a large number of serial sex offenders, including where an untested acquaintance rape kit matched with evidence from an unsolved stranger rape kit. Ayres et al.

(2017) provide a theoretical understanding of why the number of repeat offenders is underestimated, particularly for underreported crimes. While it may be that these behaviors are widespread purely because of a lack of enforcement, it seems likely that many offenders simply are not deterred by the potential punishment, perhaps because they gain too much through their offenses or because they do not see their behavior as illegal.

Cheng & Hsiaw (2022) study a model with a continuum of perpetrator types as well as a continuum of reporting agent experiences using a global games framework. Agents are intrinsically motivated to report misconduct, but also pay costs for reporting, which are increased if the report is not followed by a sanction. Applying this framework to an information escrow, they find that whether an escrow increases or decreases reporting depends on how much utility agents gain from filing an unreleased report in the escrow.

In this paper, we focus on improving the design of an escrow system, rather than on the comparison of reporting with and without an escrow. Callisto and the Reporting Lockbox have been implemented, and other proposals exist for extending the use of these escrows: Ayres (2018) considers an escrow with a governmental authority to prevent repeat offenders from hiding behind nondisclosure agreements; Hemel & Lund (2018) considers using state human rights agency as information escrow for workplace sexual misconduct allegations. In addition, some experimental evidence suggests that escrows do increase reporting, with Ayres (2017) involving allegation escrows and Babcock & Landeo (2004) involving a settlement escrow based on Gertner & Miller (1995). Besides this paper, other work on improving escrow systems generally has focused on improving the security of information stored in escrow. (Arun et al., 2018; Rajan et al., 2018)

Although not examining information escrows, Chassang & Miquel (2019) use mechanism design to consider a single whistleblower whose reporting cost stems from potential retaliation. Along with subsequent work in Chassang & Zehnder (2019) and Boudreau et al. (2023), they consider introducing garbling into reporting systems to reduce retaliation and encourage reporting. Their approach removes the coordination issues key to our approach.

3.2 Model Basics

We model a firm designing a system for its employees to report misconduct by a manager, with the goal of minimizing that misconduct. We consider two periods. At

the end of the first period, the firm receives reports from the employees about whether they experienced misconduct. Based on those reports, the firm updates its belief about how likely the manager is to engage in misconduct and then decides whether to retain or fire the manager. In the second period, the retained first-period manager or newly hired second-period manager interacts with the employees, leading to the possibility of more misconduct.

Sequence and Information

In more detail, a manager and n employees, $i \in \{1, 2, \ldots, n\}$ work for the same firm. The manager has type μ, which everyone knows was drawn from a pool with distribution $f(\mu)$.

In the first of two periods, all employees interact with the manager, resulting in

$$x_{1i} = \begin{cases} 1, & \text{misconduct with probability } \mu \\ 0, & \text{otherwise.} \end{cases}$$

The employees report to the firm:

$$r_i = \begin{cases} 1, & \text{a costly report recommending firing} \\ 0, & \text{a costless report recommending keeping.} \end{cases}$$

Here, we assume an asymmetry in reporting costs: it is costly to recommend firing (i.e., to report misconduct) but costless to recommend keeping (i.e., to report no misconduct). This direction of asymmetry captures some intuitions about situations in which misconduct reporting arise. The manager is hired and changing that requires some active effort to get the manager fired, while the report of no misconduct can be costless silence. In addition, detailing and reliving the experience of misconduct may be inherently costly.[3]

Then, the firm chooses whether to terminate the manager,

$$t = \begin{cases} 0, & \text{keep manager} \\ 1, & \text{fire manager.} \end{cases}$$

[3]Note that in situations with a manager who is already being fired, but where the firm might be swayed to change its decision by reports of the manager's good behavior, the active effort cost goes in the opposite direction (making reporting no misconduct more costly than reporting misconduct). However, if there are costs inherent in reporting misconduct, like the detailing and reliving of the experience, those costs could outweigh the cost of making a report advocating for changing the status quo. This paper covers the latter situation, where reporting misconduct is always the more costly report, even when the manager has a higher type than the pool, making the firm predisposed to fire the manager without any additional information.

where upon firing the first manager, the firm must hire a new manager drawn from a pool with distribution $g(\gamma)$.

In the second period, all employees interact with the second-period manager (who may be the same manager kept from the first period), resulting in

$$x_{2i} = \begin{cases} 1, & \text{misconduct with probability equal to manager's type} \\ 0, & \text{otherwise.} \end{cases}$$

Utilities

The firm minimizes misconduct:

$$u_{Firm} = \sum_i (-x_{1i} - x_{2i})$$

The employees maximize their own experiences but pay costs c_r for the costly report:

$$u_i = -x_{1i} - x_{2i} - c_r r_i$$

As noted above, we are not taking the manager to be a strategic actor.

3.3 First-Best: Firm Observes the First Period

To give a baseline for the mechanism design problem that follows, we first consider a firm with full information about the manager's first-period behavior. If the firm can observe the first-period manager directly and sees that m of the n interactions are misconduct, then the firm can update on the first-period manager's type. The posterior distribution and expectation of μ will be

$$\begin{aligned} f(\mu|m \text{ misconduct}) &= \frac{\Pr(m|\mu) f(\mu)}{\int_0^1 \Pr(m|\mu') f(\mu') d\mu'} \\ &= \frac{\mu^m (1-\mu)^{n-m} f(\mu)}{\int_0^1 \mu'^m (1-\mu')^{n-m} f(\mu') d\mu'} \\ E(\mu|m) &= \int_0^1 \mu f(\mu|m) d\mu \end{aligned}$$

A new manager would be drawn from the distribution $g(\gamma)$, and so the expectation of the type of a new manager is

$$E(\gamma) = \int_0^1 \gamma g(\gamma) d\gamma$$

Then, as the firm is simply minimizing misconduct,

$$E(u_{Firm}(\text{keep})) = -m - nE(\mu|m)$$
$$E(u_{Firm}(\text{fire})) = -m - nE(\gamma)$$

and the firm will fire the first manager if

$$-nE(\gamma) > -nE(\mu|m)$$
$$E(\mu|m) > E(\gamma)$$

Intuitively, the firm will fire the first manager if its posterior belief about the manager's type gives a higher probability of misconduct than a random draw from the hiring pool.

3.4 Model Version 1: No Reporting Cost; Beta Distributions as Priors

For the remainder of this paper, we will use beta distributions as the priors for the manager and the hiring pool. This simplification allows greater tractability, and we can think of the managers as parallel to biased coins, where each employee interaction is like a coin flip where the probability of heads/misconduct is the manager's type. In this section, we also simplify the model by removing costly reporting.

Let the prior for the type of the pool of managers be the beta distribution $\text{Beta}(\alpha_0, \beta_0)$, with $\alpha_0, \beta_0 > 0$. Then the expectation of the type for a newly drawn manager is

$$\bar{\gamma} \equiv \frac{\alpha_0}{\alpha_0 + \beta_0}$$

We can imagine the first-period manager coming from this same pool, but with additional information leading to an updated prior. To be specific, after the first-period manager is hired, but before the individual interactions with employees take place, the manager sends a public signal of their type through their public behavior that causes everyone to update the distribution of the first-period manager's type to $\text{Beta}(\alpha, \beta)$, with $\alpha, \beta > 0$.

First-Best

Here, we evaluate the first-best firing condition from above,

$$E(\mu|m) > E(\gamma),$$

under these beta distribution priors. If the firm sees m incidents of misconduct out of n interactions, the posterior on the first-period manager's type would be $\text{Beta}(\alpha + m, \beta + n - m)$, and so the firm would fire if

$$\frac{\alpha + m}{\alpha + \beta + n} > \bar{\gamma}$$
$$m > \bar{\gamma}(\alpha + \beta + n) - \alpha$$

A Single-Threshold Mechanism with Costless Reporting and Full Commitment Achieves the First-Best

Next we consider the firm committing to a mapping $\tau : \{0, 1\}^n \to \{0, 1\}$ from reports r_i to a decision about whether to terminate the manager t. We restrict attention to truthful, pure elicitation mechanisms that are monotonic (with τ weakly increasing in reports of misconduct) and anonymous (with all reports being treated the same).

Consider a threshold number of reports for firing: $\bar{m} \equiv \lfloor \bar{\gamma}(\alpha+\beta+n)-\alpha \rfloor$. Note that if $\bar{m} < 0$, the firm will always fire the first manager regardless of whatever misconduct occurs at the firm, and if $\bar{m} \geq n$, the firm will never fire the first manager. The firm can achieve the first-best without any information from the employees.

Otherwise, we will show that the firm can achieve the first-best by setting a threshold number of reports for firing at $\bar{m}$. In other words, the firm only fires if the number of reports strictly exceeds $\bar{m}$.

Note that the firm and the employees have completely aligned interests. Both only want to fire the first manager if the posterior expectation of the manager's type is worse than a random draw from the hiring pool.

Each employee i knows that their report r_i will affect the firm's decision only if exactly $\bar{m}$ reports have also been filed. In this situation, the employee is *pivotal* to the decision. Whenever the employee is not pivotal, the employee's choice to report misconduct or not will have no effect on the firing decision and the employee's utility, so the employee only considers the benefit of reporting misconduct conditional on their own pivotality.

Truthful reporting (i.e., recommending firing only upon experiencing misconduct) will be a Bayesian Nash equilibrium (BNE). An employee will report misconduct if

and only if

$$\begin{aligned}
\Pr(\text{pivotal}|x_{1i})[\text{net benefit of reporting misconduct vs. not}|\text{pivotal}, x_{1i}] &\geq 0 \\
[\text{net benefit of reporting misconduct vs. not}|\text{pivotal}, x_{1i}] &\geq 0 \\
-\bar{\gamma} - (-E[\mu|\text{pivotal}, x_{1i}]) &\geq 0 \\
E[\mu|\text{pivotal}, x_{1i}] - \bar{\gamma} &\geq 0
\end{aligned}$$

Here, if $x_{1i} = 0$, there will be $\bar{m}$ total misconduct interactions, and the employee will prefer not to report, and if $x_{1i} = 1$, there will be $\bar{m} + 1$ total misconduct interactions, and the employee will prefer to report. Because the firm and the employee have the same preferences over managers, the firm using its own first-best threshold for firing also incentivizes truthful reporting.

Proposition 4. *When reports are costless, a single-threshold mechanism with a threshold of $\bar{m}$ will elicit truthful reports and will be optimal, allowing the firm to achieve its first-best.*

Note that this set-up with costless reporting closely parallels the Condorcet Jury Theorem-related analysis in Austen-Smith & Banks (1996). There, jurors receive signals that are independently drawn from a state-dependent distribution for states A and B, and then vote for either A or B after observing their individual signal and having a belief about the prior probability that the state is A. An aggregation rule sets a threshold k_f such that B is the outcome if and only if more than k_f votes are for B. Jurors only care about their vote conditional on being pivotal. Austen-Smith & Banks (1996) find that a juror voting according to their own signal is rational if and only if the vote threshold k_f equals the threshold value for the sum of the signals (k^*) above which the posterior probability that the state is B is greater than a half. Here, the interactions with the manager are independent draws that depend on manager type, the aggregation rule is the threshold above which the firm fires the manager, and a voting threshold of $\bar{m}$ equals the threshold for the sum of the interactions x_{1i} above which a new hire from the pool is preferred to the first-period manager.

3.5 Model Version 2: Reporting Cost c_r; Beta Distributions as Priors

We now consider the impact of having one kind of report be costly, instead of having both kinds be costless. As before, the expectation of the type for a newly drawn manager is $\bar{\gamma}$, and the distribution of the first-period manager's type is Beta(α, β).

The first-best for the firm remains the same. We again consider single-threshold mechanisms.

We assume here that reporting misconduct costs the employees $c_r \geq 0$ and that the firm is setting some single threshold m such that they fire the first-period manager if and only if the number of reports strictly exceeds m. Again, when $\bar{m} < 0$ or $\geq n$, the firm can achieve the first-best without eliciting information from the employees. We try to find a BNE with truthful reporting for $\bar{m} \in [0, n-1]$, in other words for parameter values such that information from employees can affect the firing decision. So, we are searching for the existence of a truthful elicitation mechanism with a single threshold determining the firing decision.

Again, each employee knows that they will be pivotal if and only if m other reports of misconduct have been made. Since reporting misconduct now is costly, an employee i will report misconduct if and only if

$$\Pr(\text{pivotal}|x_{1i})\,(E[\mu|\text{pivotal}, x_{1i}] - \bar{\gamma}) \geq c_r$$

So, truthful reporting requires reporting misconduct after experiencing it

$$\Pr(\text{pivotal}|x_{1i} = 1)\,(E[\mu|\text{pivotal}, x_{1i} = 1] - \bar{\gamma}) \geq c_r$$

and reporting no misconduct after not experiencing it

$$\Pr(\text{pivotal}|x_{1i} = 0)\,(E[\mu|\text{pivotal}, x_{1i} = 0] - \bar{\gamma}) \leq c_r$$

Since we are looking for a BNE in which all employees report truthfully, i assumes that the other employees report truthfully, and knows that they will be pivotal if and only if m of the other employees experienced misconduct. Thus,

$$\Pr(\text{pivotal}|x_{1i}) = \binom{n-1}{m} \int_0^1 \mu^m (1-\mu)^{(n-1-m)} f(\mu|x_{1i}) d\mu$$

Since $f(\mu) = \text{Beta}(\alpha, \beta)$,

$$\begin{aligned} f(\mu|x_{1i}) &= \text{Beta}(\alpha + x_{1i}, \beta + 1 - x_{1i}) \\ &= \frac{1}{\mathcal{B}(\alpha + x_{1i}, \beta + 1 - x_{1i})} \mu^{\alpha + x_{1i} - 1} (1-\mu)^{\beta - x_{1i}} \end{aligned}$$

Substituting in,

$$\begin{aligned} \Pr(\text{pivotal}|x_{1i}) &= \binom{n-1}{m} \int_0^1 \mu^m (1-\mu)^{(n-1-m)} \frac{1}{\mathcal{B}(\alpha + x_{1i}, \beta + 1 - x_{1i})} \mu^{\alpha + x_{1i} - 1} (1-\mu)^{\beta - x_{1i}} d\mu \\ &= \binom{n-1}{m} \frac{1}{\mathcal{B}(\alpha + x_{1i}, \beta + 1 - x_{1i})} \int_0^1 \mu^{m + \alpha + x_{1i} - 1} (1-\mu)^{n-1-m+\beta - x_{1i}} d\mu \end{aligned}$$

Then, since $\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \mathcal{B}(a,b)$,

$$\Pr(\text{pivotal}|x_{1i}) = \binom{n-1}{m}\frac{\mathcal{B}(m+\alpha+x_{1i}, n-m+\beta-x_{1i})}{\mathcal{B}(\alpha+x_{1i}, \beta+1-x_{1i})}$$

In addition, we know that

$$\begin{aligned} E[\mu|\text{pivotal}, x_{1i}] &= \frac{\alpha+m+x_{1i}}{\alpha+m+x_{1i}+\beta+n-1-m+(1-x_{1i})} \\ &= \frac{\alpha+m+x_{1i}}{\alpha+\beta+n} \end{aligned}$$

Now, the two conditions for truthful reporting[4] become

$$\begin{aligned} NBT \equiv \Pr(\text{pivotal}|x_{1i}=1)\,(E[\mu|\text{pivotal}, x_{1i}=1] - \bar{\gamma}) &\geq c_r \\ \binom{n-1}{m}\frac{\mathcal{B}(m+\alpha+1, n-m+\beta-1)}{\mathcal{B}(\alpha+1,\beta)}\left(\frac{\alpha+m+1}{\alpha+\beta+n} - \bar{\gamma}\right) &\geq c_r \end{aligned}$$

and

$$\begin{aligned} NBF \equiv \Pr(\text{pivotal}|x_{1i}=0)\,(E[\mu|\text{pivotal}, x_{1i}=0] - \bar{\gamma}) &\leq c_r \\ \binom{n-1}{m}\frac{\mathcal{B}(m+\alpha, n-m+\beta)}{\mathcal{B}(\alpha,\beta+1)}\left(\frac{\alpha+m}{\alpha+\beta+n} - \bar{\gamma}\right) &\leq c_r. \end{aligned}$$

We are seeking pairs of thresholds and reporting costs (m, c_r) such that both conditions are satisfied. We also introduce notation above, where NBT is the net benefit of truthfully reporting misconduct and NBF is the net benefit of falsely reporting misconduct.

Figure 3.1 gives an example of a plot of NBT and NBF as functions of m. We consider each pair of dots with the same m value. A particular threshold m can be part of an (m, c_r) pair with truthful reporting when the $NBT(m) \geq NBF(m)$, and the c_r values that will support truthful reporting will be all of the values in between the curves: $c_r \in [NBF(m), NBT(m)]$. Alternatively, if we start with some c_r we

[4]The concept behind this unwillingness to report is different from the usual logic. Usually, we think that someone may not be willing to report because they worry that the firm will not take action and they will pay the cost of reporting and not get the result they want. Here, the firm will fire upon the pivotal report, but it may still be undesirable for the employee to report because the cost of reporting is greater than the net benefit from the firing (even though in a costless world the employee prefers a new draw to the first-period manager). Note that we need to justify the cost c_r as something that is unconditional on the result, for example the costliness of making the report and having to relive/recount the experience in detail.

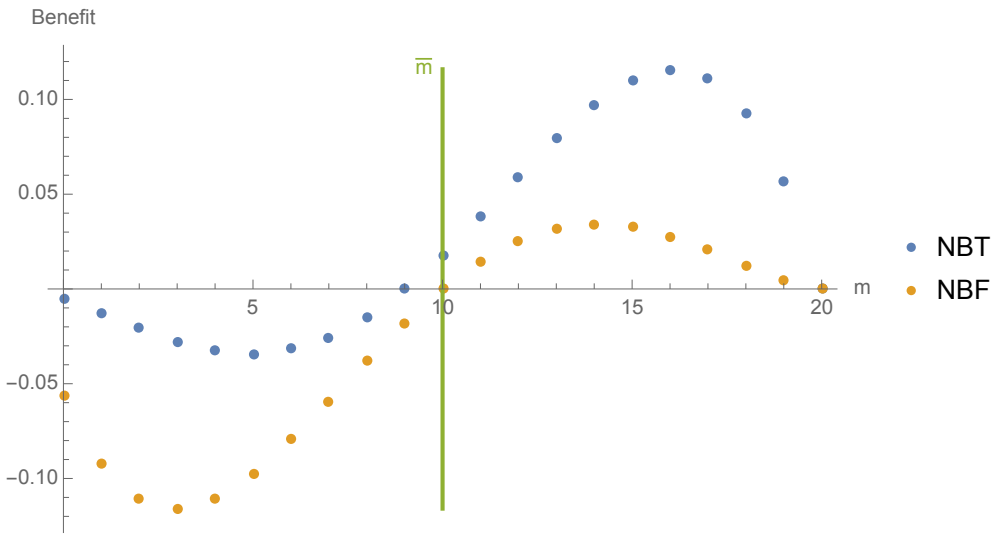


Figure 3.1: $\alpha = 2, \beta = 2, n = 20, \bar{\gamma} = 0.5$

can find all thresholds m that will support truthful reporting by seeing where the horizontal line at c_r lies between the NBT and NBF curves with NBT on top.

From the prior version of the model with no reporting costs, we know that these conditions will be satisfied for $c_r = 0$ and $m = \bar{m}$. Note that $\bar{m}$ is the only value of m that will satisfy these conditions for $c_r = 0$, because increasing m to $\bar{m} + 1$ or any higher value of m must violate the second condition, as every term on the left-hand side will be positive. In addition, this first-best result can be achieved for any

$$c_r \leq \binom{n-1}{\bar{m}} \frac{\mathcal{B}(\bar{m}+\alpha+1, n-\bar{m}+\beta-1)}{\mathcal{B}(\alpha+1,\beta)} \left(\frac{\alpha+\bar{m}+1}{\alpha+\beta+n} - \bar{\gamma} \right)$$

using the threshold $m = \bar{m}$.

Corollary 1. *The interval* $[NBF(m), NBT(m)]$ *will always exist at* $\bar{m}$*, and all reporting costs in that interval will support truthful reporting and the firm's first-best outcome. No other interval* $[NBF(m), NBT(m)]$ *will include 0.*

We can see one example of this interval at the vertical line in Figure 3.1.

Note than any $m < \bar{m}$ will lead to $E[\mu|\text{pivotal}, x_{1i}] - \bar{\gamma} < 0$. Since we assume that the reporting cost c_r cannot be negative, there will be no truthful reporting possible for these lower thresholds. This conclusion squares with the intuition that a lower number of misconduct incidents makes the employee prefer retaining the manager,

even without reporting costs, and the addition of reporting costs will not make firing the manager more desirable.

Next, consider thresholds $m > \bar{m}$. Will there always be some c_r such that truthful reporting is incentive-compatible? The necessary condition is that the interval described above exists, meaning that

$$\binom{n-1}{m}\frac{\mathcal{B}(m+\alpha+1, n-m+\beta-1)}{\mathcal{B}(\alpha+1,\beta)}\left(\frac{\alpha+m+1}{\alpha+\beta+n}-\bar{\gamma}\right) \geq$$
$$\binom{n-1}{m}\frac{\mathcal{B}(m+\alpha, n-m+\beta)}{\mathcal{B}(\alpha,\beta+1)}\left(\frac{\alpha+m}{\alpha+\beta+n}-\bar{\gamma}\right)$$

Using the identities $\mathcal{B}(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and $\Gamma(a+1) = a\Gamma(a)$,

$$\binom{n-1}{m}\frac{\mathcal{B}(m+\alpha, n-m+\beta-1)}{\mathcal{B}(\alpha,\beta)}\left(\frac{m+\alpha}{\alpha}\right)\left(\frac{\alpha+m+1}{\alpha+\beta+n}-\bar{\gamma}\right) \geq$$
$$\binom{n-1}{m}\frac{\mathcal{B}(m+\alpha, n-m+\beta-1)}{\mathcal{B}(\alpha,\beta)}\left(\frac{n-m+\beta-1}{\beta}\right)\left(\frac{\alpha+m}{\alpha+\beta+n}-\bar{\gamma}\right)$$

Because $\binom{n-1}{m}\frac{\mathcal{B}(m+\alpha,n-m+\beta-1)}{\mathcal{B}(\alpha,\beta)}$ is positive, the condition above will hold if and only if

$$\left(\frac{m+\alpha}{\alpha}\right)\left(\frac{\alpha+m+1}{\alpha+\beta+n}-\bar{\gamma}\right) \geq$$
$$\left(\frac{n-m+\beta-1}{\beta}\right)\left(\frac{\alpha+m}{\alpha+\beta+n}-\bar{\gamma}\right)$$

Remember that at $m = \bar{m}$, the left-hand side is positive and the right-hand side is negative. Note that when $m > \bar{m}$ all four terms must be positive, given that $m \leq n-1$.

Then, an interval exists when

$$\left(\frac{m+\alpha}{\alpha}\right)\left(\frac{\alpha+m+1}{\alpha+\beta+n}-\bar{\gamma}\right)-\left(\frac{n-m+\beta-1}{\beta}\right)\left(\frac{\alpha+m}{\alpha+\beta+n}-\bar{\gamma}\right) \geq 0 \qquad (3.1)$$

which is a convex quadratic equation in m.

Note that the condition in (3.1) will always hold if

$$\frac{(m+\alpha)\beta}{(n-m+\beta-1)\alpha} \geq 1$$
$$m \geq \frac{\alpha}{\alpha+\beta}(n-1)$$

and since $\frac{\alpha}{\alpha+\beta}(n-1) < n-1$, there will always exist an interval of the highest values of m that satisfies (3.1).

Proposition 5. *The interval* $[NBF(m), NBT(m)]$ *will always exist for an upper range of thresholds that includes the highest threshold,* $m = n - 1$*. Reporting costs in* $[NBF(m), NBT(m)]$ *for an* m *in this range will support truthful reporting.*

Thus, there always exist two ranges of values for m such that values of c_r permitting truthful reporting exist: 1) a range with $\bar{m}$ as its lowest value and 2) a range with $n - 1$ as its highest value.

Gaps Can Exist in the Ranges of Thresholds and Reporting Costs that Support Truthful Reporting

However, it is possible for intermediate values of m to violate (3.1); this violation is possible when the left-hand side of (3.1) has two zeroes. The discriminant is a convex quadratic function of $\bar{\gamma}$:

$$(\alpha + \alpha^2 + \beta + \alpha\beta - \alpha^2\bar{\gamma} - 2\alpha\beta\bar{\gamma} - \beta^2\bar{\gamma} - \alpha n - \alpha n\bar{\gamma} - \beta n\bar{\gamma})^2 -$$
$$4(\alpha + \beta)(\alpha^2 + \alpha\beta - \alpha^2\bar{\gamma} - \alpha\beta\bar{\gamma} - \alpha^2 n - \alpha n\bar{\gamma} + \alpha^2 n\bar{\gamma} + \alpha\beta n\bar{\gamma} + \alpha n^2\bar{\gamma})$$

Note that when the prior on the first manager's type equals that of the pool ($\frac{\alpha}{\alpha+\beta} = \bar{\gamma}$), this discriminant simplifies to

$$\alpha^2 - 4\alpha^2\beta + \beta^2 - 4\alpha\beta^2 + 2\alpha\beta - 4\alpha\beta n$$

which is always negative if $\alpha, \beta > 0.25$.[5] Thus, for $\alpha, \beta > 0.25$, if the first manager looks like the pool of replacements prior to interacting with employees, there will be no threshold values $m > \bar{m}$ without a range of values for c_r that permit truthful reporting.

When $\bar{\gamma} \to 0$, the discriminant approaches

$$\begin{aligned}&(\alpha + \alpha^2 + \beta + \alpha\beta - \alpha n)^2 - 4(\alpha + \beta)(\alpha^2 + \alpha\beta - \alpha^2 n)\\ =&[(\alpha + \beta)(\alpha - 1) + \alpha n]^2\end{aligned}$$

and when $\bar{\gamma} \to 1$, it approaches

$$\begin{aligned}&(\alpha + \beta - \alpha\beta - \beta^2 - 2\alpha n - \beta n)^2 - 4(\alpha + \beta)(\alpha n^2 + \alpha\beta n - \alpha n)\\ =&[(\alpha + \beta)(n + \beta - 1) - \alpha n]^2\end{aligned}$$

so the discriminant will always be positive for extreme values of $\bar{\gamma}$, meaning that there is an interval of values of $\bar{\gamma} \in (0, 1)$ that includes $\frac{\alpha}{\alpha+\beta}$ such that (3.1) is satisfied

[5] This restriction excludes some bimodal distributions, with high probabilities of being either very unlikely to cause misconduct or very likely to do so.

for all thresholds, but outside of that interval, (3.1) is unsatisfied for some values of m. Equivalently, if $\alpha, \beta > 0.25$, when the first-period manager is similar enough to the replacement pool, it will be possible to find a reporting cost permitting truthful reporting for every threshold m, but when they are sufficiently different from the pool, there may be values of m such that no reporting cost exists that makes truthful reporting incentive compatible.

Example 1. *When the manager has a worse type than the pool:* $\frac{\alpha}{\alpha+\beta} > \bar{\gamma}$*, it is possible for* $NBT < NBF$ *for some* $m \in (\bar{m}, n-1)$.

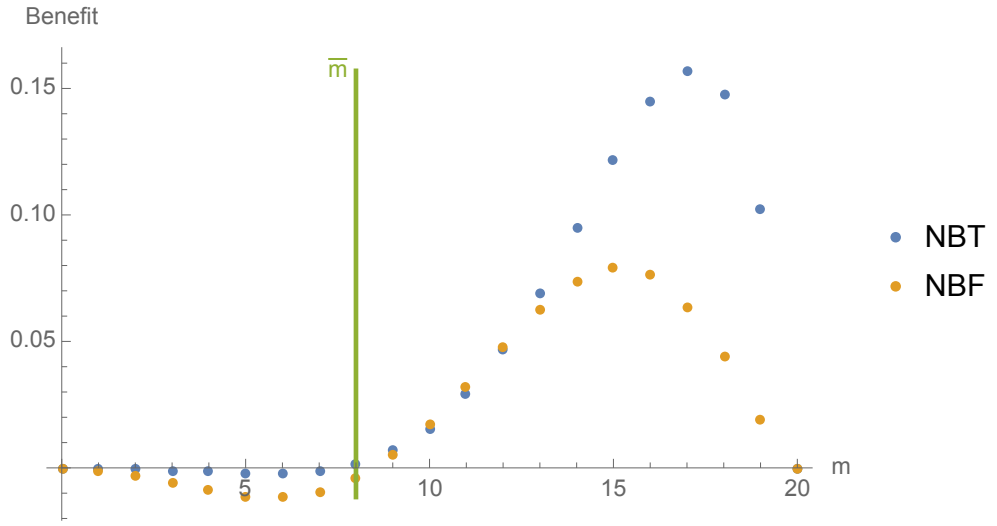


Figure 3.2: $\alpha = 5, \beta = 2, n = 20, \bar{\gamma} = 0.5$

Consider the example in Figure 3.2, where $\alpha = 5, \beta = 2, n = 20$, and $\bar{\gamma} = 0.5$. Here, $NBT > NBF$ when $m = \bar{m} = 8$, when $m = 9$, and when $m \in [13, 19]$. However, $NBF > NBT$ when m equals 10, 11, or 12. The curves flip, and the thresholds that can support truthful reporting are not a set of consecutive numbers.

Why does this flip occur? Recall that

$$NBT \equiv \Pr(\text{pivotal}|x_{1i} = 1)\,(E[\mu|\text{pivotal}, x_{1i} = 1] - \bar{\gamma})$$
$$\binom{n-1}{m} \frac{\mathcal{B}(m+\alpha+1, n-m+\beta-1)}{\mathcal{B}(\alpha+1, \beta)} \left(\frac{\alpha+m+1}{\alpha+\beta+n} - \bar{\gamma}\right)$$

and

$$NBF \equiv \Pr(\text{pivotal}|x_{1i}=0)\,(E[\mu|\text{pivotal},x_{1i}=0]-\bar{\gamma})$$
$$\binom{n-1}{m}\frac{\mathcal{B}(m+\alpha,n-m+\beta)}{\mathcal{B}(\alpha,\beta+1)}\left(\frac{\alpha+m}{\alpha+\beta+n}-\bar{\gamma}\right).$$

Note that for an employee who experiences misconduct, the difference between the manager's expected type and the pool is $\left(\frac{\alpha+m+1}{\alpha+\beta+n}-\bar{\gamma}\right)$, while it is $\left(\frac{\alpha+m}{\alpha+\beta+n}-\bar{\gamma}\right)$ for an employee who does not experience misconduct. As in Figure 3.3, these are linear functions of m with the same slope, where the line for an employee experiencing misconduct is a constant $\left(\frac{1}{\alpha+\beta+n}\right)$ higher.

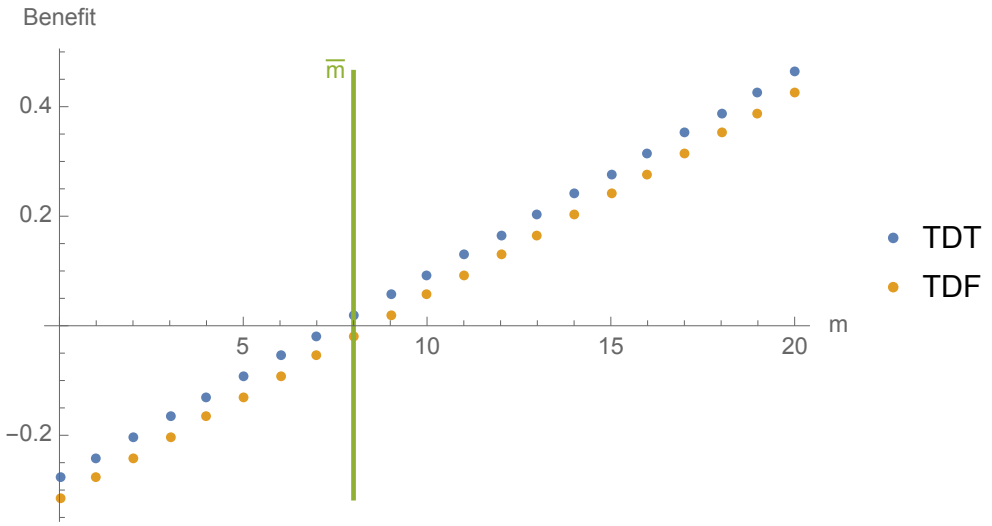


Figure 3.3: $\alpha = 5$, $\beta = 2$, $n = 20$, $\bar{\gamma} = 0.5$; expected type difference between manager and pool.

The flip then occurs entirely because of differences in the probabilities that the employee will be pivotal. These pivot probabilities for an employee experiencing misconduct and one who does not are shown in Figure 3.4.

Here, the chance that an employee who does not experience misconduct will be pivotal is greater than the chance for an employee who experiences misconduct, until the threshold gets sufficiently high (at $m = 14$). For the given parameter values, the prior expectation of manager's type $\left(\frac{5}{7}\right)$ is worse than the hiring pool $\left(\frac{1}{2}\right)$. An employee who experiences misconduct is less likely to have a misconduct report affect the firing decision because this bad prior about the manager leads them to think that it is more likely that enough others will report misconduct without

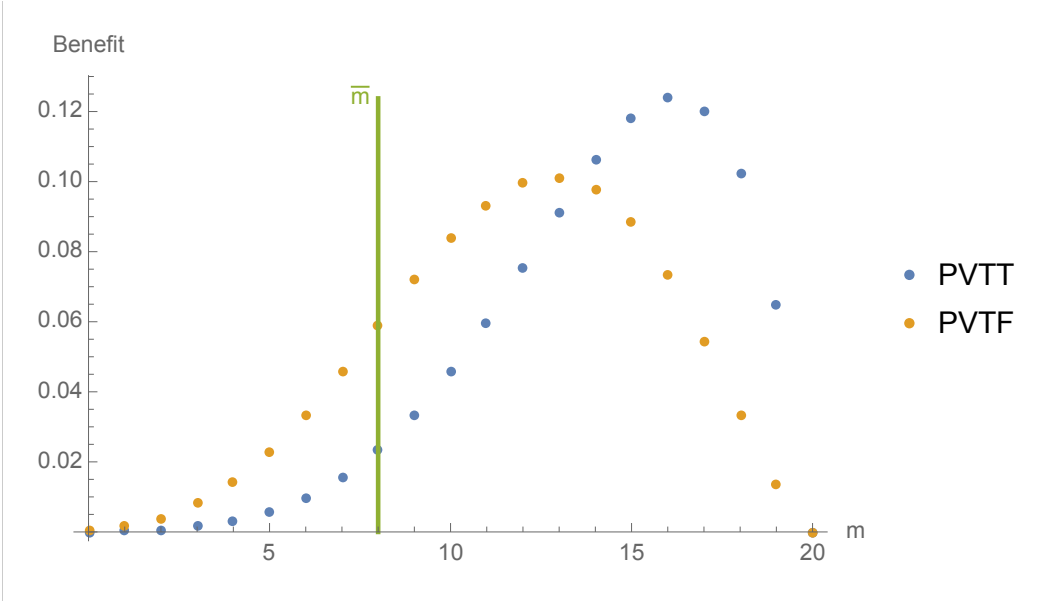


Figure 3.4: $\alpha = 5$, $\beta = 2$, $n = 20$, $\bar{\gamma} = 0.5$; pivot probabilities.

them than an employee who does not experience misconduct would think. In essence, the employees who experience misconduct know that the manager started out bad and now (because of experiencing misconduct) think that the manager is even worse, and so are more likely to believe that they can free ride off of others' reports of misconducts and gain the benefits of firing without having to pay the cost of reporting.[6]

Note also that the employees only care about the pivot probabilities because of costly reporting. If reporting is costless, the employees only care if their expected utility is positive or negative, and changes in the pivot probability term will not change the sign.

Example 2. *When the manager has a worse type than the pool:* $\frac{\alpha}{\alpha+\beta} > \bar{\gamma}$*, it is possible for the reporting costs* c_r *that support truthful reporting not to form a continuous interval.*

The parameters in the earlier Figure 3.2 lead to a such a gap in reporting costs, and the example in Figure 3.5 shows a larger, clearer gap. If $\alpha = 10$, $\beta = 2$, $n = 20$, and $\bar{\gamma} = 0.5$, then setting a threshold of $\bar{m} = 6$ will yield truthful reporting

[6]Numerical analysis suggests that this kind of flip does not occur for $m > \bar{m}$ when the prior expectation of the manager's type is better than the hiring pool. Flips do occur for thresholds below $\bar{m}$, but those thresholds can never support truthful reporting because NBT and NBF are both negative.

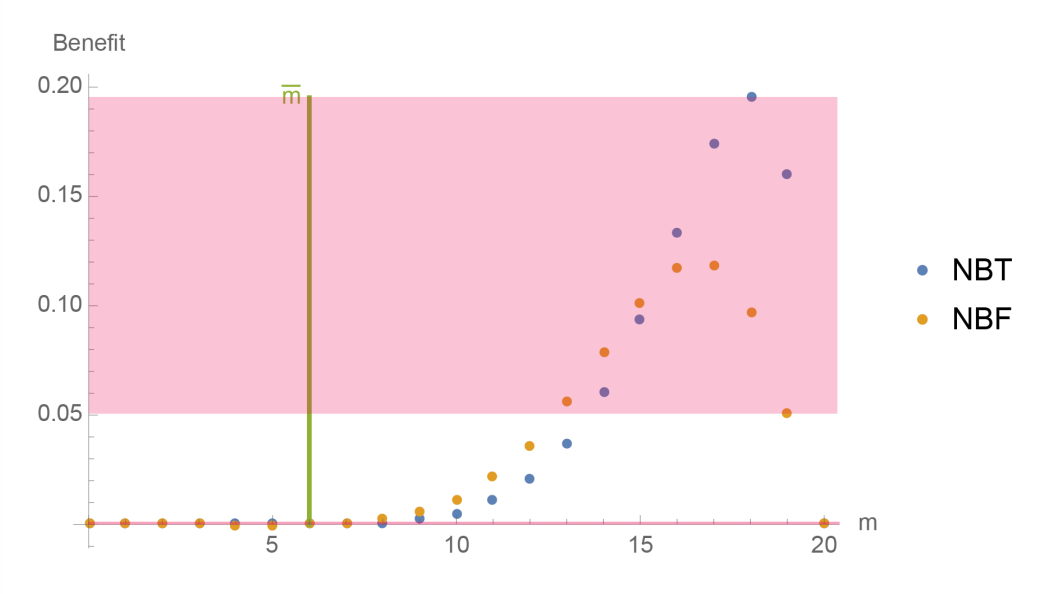


Figure 3.5: $\alpha = 10$, $\beta = 2$, $n = 20$, $\bar{\gamma} = 0.5$; two pink regions show reporting costs for which truthful reporting is possible under some threshold.

for $c_r \in [0, 9 \times 10^{-4}]$. This region is the sliver of pink at the bottom of Figure 3.5. However, thresholds between 7 and 15 do not support truthful reporting, and thresholds of 16, 17, 18, and 19 support truthful reporting for $c_r \in [0.12, 0.13]$, $[0.12, 0.17]$, $[0.10, 0.20]$, and $[0.05, 0.16]$ respectively, spanning the upper pink region. There is a gap in the range of reporting costs for which truthful reporting is possible, where any $c_r \in (9 \times 10^{-4}, 0.05)$ cannot lead to truthful reporting in a single-threshold mechanism.

The gap in thresholds that support truthful reporting leads to this gap in reporting costs that support truthful reporting.[7] It means that there may be reporting costs for which no truthful, pure elicitation, single-threshold mechanism exists, and not solely for the case when the reporting cost gets too high.

Firm's Decision to Elicit Information through a Single-Threshold Mechanism

In the previous sections, we have considered whether a truthful single-threshold mechanism exists. Now we turn to thinking about the designer's decisions. Should the firm elicit information from the employees at all?

Without any information from misconduct reports, the firm will choose purely on

[7]There is also the potential for gaps in reporting costs due to integer effects, where, even though the curves for NBT and NBF would span a continuous range, the intersection of the ranges at integer values of m would not.

the basis of its prior beliefs about the first-period manager and the pool, and will fire the first manager if

$$E(\mu) > E(\gamma)$$
$$\frac{\alpha}{\alpha+\beta} > \bar{\gamma}$$

Proposition 6. *If any truthful single-threshold mechanism exists and if the manager is better (has a lower type) than the pool* $\left(\frac{\alpha}{\alpha+\beta} < \bar{\gamma}\right)$*, the firm always benefits from using the mechanism.*

When the prior on the manager is better than the pool and the firm gathers no additional information, the firm will always retain the first-period manager. Here, any reporting is an improvement, because even if the threshold is much higher than the firm's ideal threshold $\bar{m}$, the result will always be to fire managers who engage in enough misconduct to make the posterior on their type worse than the pool. Any truthful single-threshold mechanism therefore is expected to get rid of only bad actors who would have been retained without eliciting information.

Proposition 7. *If a truthful single-threshold mechanism exists with threshold* m_M *and if the manager is worse (has a higher type) than the pool* $\left(\frac{\alpha}{\alpha+\beta} > \bar{\gamma}\right)$*, the firm benefits from using the mechanism when*

$$\sum_{m=0}^{m_M} \binom{n}{m} \frac{\mathcal{B}(m+\alpha, n-m+\beta)}{\mathcal{B}(\alpha,\beta)} \left(\bar{\gamma} - \frac{\alpha+m}{\alpha+\beta+n}\right) > 0$$

Here, since $E(\mu) > E(\gamma)$, without reporting the firm always would fire the manager. Ideally, the firm would like to observe the episodes of misconduct and fire only when there are $(\bar{m}+1)$ or more such episodes. A single-threshold mechanism may instead set a threshold $m_M > \bar{m}$. Then, under either the prior or the mechanism, (m_M+1) or more misconduct interactions would lead to firing. However, for 0 to m_M misconduct interactions, the mechanism would lead to a different outcome (retaining the manager), and the firm only prefers to retain a manager with $\bar{m}$ or fewer misconduct interactions. Thus, the firm prefers the mechanism outcome only for the lower values of m as illustrated in Figure 3.6.

The difference between the firm's expected utility under the mechanism and under the prior is the probability that m misconduct interactions occur (given that Beta(α,β) is the prior on the manager's type) multiplied by the difference between the manager

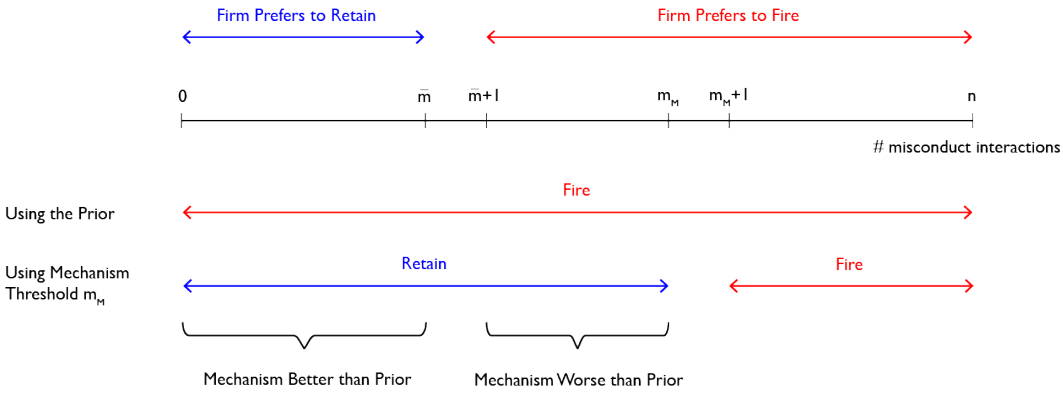


Figure 3.6: Comparing firm's firing decision under the prior and a single-threshold mechanism when the prior is worse than the pool.

and pool's expected types, summed over the values of m where the mechanism retains a manager who would have been fired under the prior:

$$\sum_{m=0}^{m_M} Pr(m)\left[-E(\mu|m) - (-\bar{\gamma})\right] = \sum_{m=0}^{m_M} \binom{n}{m} \frac{\mathcal{B}(m+\alpha, n-m+\beta)}{\mathcal{B}(\alpha,\beta)} \left(\bar{\gamma} - \frac{\alpha+m}{\alpha+\beta+n}\right)$$

Note that $\left(\bar{\gamma} - \frac{\alpha+m}{\alpha+\beta+n}\right)$ is decreasing in m and will be negative for $m > \bar{m}$. If we focus on the region of the parameter space such that reports can make a difference, it must also be true that the manager will be better than the pool if $m = 0$. So, the firm's net expected utility from using the mechanism rather than the prior will begin positive for $m = 0$, will reach its maximum at $m = \bar{m}$, but then will decrease above $\bar{m}$, and must be negative for $m = n$.

Returning to the example in Figure 3.2, where $\alpha = 5$, $\beta = 2$, $n = 20$, and $\bar{\gamma} = 0.5$, the firm prefers a mechanism with a threshold of 10 or below, and would choose not to elicit information for a threshold of 11 or above. However, because of the flip with $NBF > NBT$ for $m \in \{10, 11, 12\}$, the only truthful mechanisms that the firm prefers to using the prior are for $\bar{m} = 8$ and 9.

For the example in Figure 3.5, where $\alpha = 10$, $\beta = 2$, $n = 20$, and $\bar{\gamma} = 0.5$, the only truthful mechanism that the firm prefers to the prior is for $\bar{m} = 6$.

Finally, note that we can extend the logic in Figure 3.6 to a comparison of different single-threshold mechanisms that support truthful reporting under the same reporting cost. The higher the threshold, the more mismatches there will be between the firm's ideal and the mechanism's output.

Corollary 2. *Given a reporting cost c_r, the firm prefers the truthful single-threshold mechanism with the lowest threshold to any other truthful single-threshold mechanisms.*

Mixing between Different Thresholds Can Improve the Firm's Outcome Compared to a Single-Threshold Mechanism

We have shown that, under costly reporting, a single-threshold mechanism can be non-optimal because the firm would prefer no information elicitation. Next, we consider whether a single-threshold mechanism can be non-optimal because the firm would prefer a mixed mechanism with multiple thresholds.

Consider the firm committing to use threshold m_1 with probability p and threshold m_2 with probability $(1 - p)$, where $m_1 < m_2$ and $m_1, m_2 \in [\bar{m}, n - 1]$. Note that if employee i is pivotal for threshold m_1, i cannot be pivotal for threshold m_2, so the new net benefit of reporting misconduct is just the sum of the net benefits of reporting misconduct for each single threshold, weighted by the probability that the threshold will be used. The conditions for truthful reporting are then

$$p[NBT(m_1)] + (1 - p)[NBT(m_2)] \geq c_r$$
$$p[NBF(m_1)] + (1 - p)[NBF(m_2)] \leq c_r$$

where $NBT(m)$ denotes the net benefit of truthfully reporting misconduct in a single-threshold mechanism with a threshold of m.

Example 3. *Mixing two thresholds can increase the firm's utility compared to the best single-threshold truthful mechanism for a given reporting cost.*

Consider a reporting cost c_r such that it is above the net benefit of truthful reporting for threshold $m > \bar{m}$, but truthful reporting is possible under cost c_r and threshold $(m + 1)$ and $(m + 1)$ is the lowest threshold supporting truthful reporting. Then, by Corollary 2, the single-threshold mechanism with threshold $(m + 1)$ is the best single-threshold mechanism for the firm. However, if c_r is also strictly above the net benefit of false reporting for threshold $(m + 1)$, the firm can mix between thresholds m and $(m + 1)$ in a proportion that makes truthful reporting incentive compatible. Furthermore, because the firm always prefers to fire managers after $(m + 1)$ reports of misconduct, the firm improves its outcome by sometimes firing those managers, instead of always retaining them under the single threshold of $(m + 1)$.

As an example, consider a reporting cost $c_r = 0.01$, when $\alpha = 2$, $\beta = 3$, $n = 20$, and $\bar{\gamma} = 0.5$.

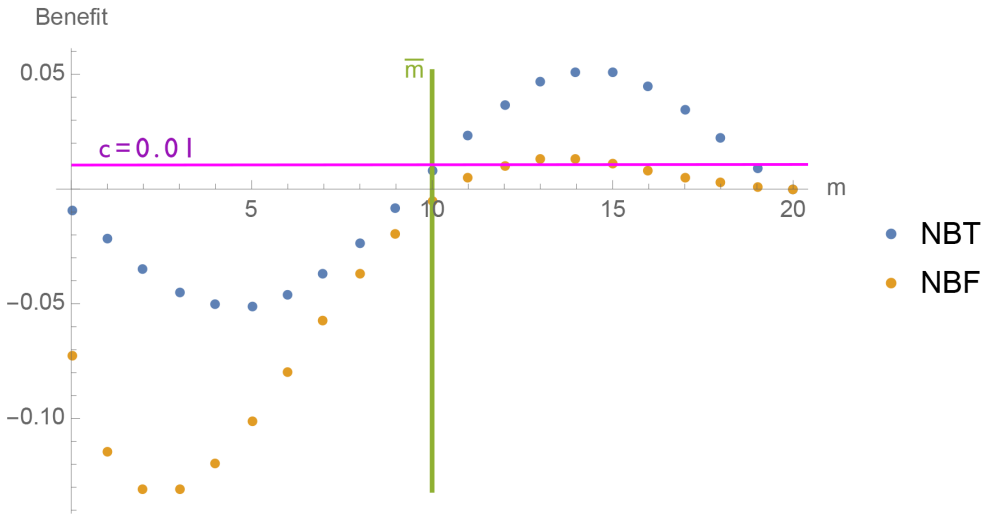


Figure 3.7: $\alpha = 2$, $\beta = 3$, $n = 20$, $\bar{\gamma} = 0.5$; reporting cost of 0.01 marked by the horizontal line.

Here, c_r is just above the range of costs that support truthful reporting under a threshold of $\bar{m} = 10$ and is within the range of costs that support truthful reporting under a threshold of 11. So, the best single-threshold truthful mechanism has a threshold of 11. Also, since the prior on the manager $\left(\frac{2}{5}\right)$ is better than the pool $\left(\frac{1}{2}\right)$, we know from Proposition 6 that this single-threshold mechanism also improves on using only the prior and always retaining the manager.

However, the firm can instead commit to using a mixture of thresholds 10 and 11. The firm can choose to make the truthful reporter indifferent between reporting misconduct and reporting no misconduct:

$$p[NBT(10)] + (1-p)[NBT(11)] = 0.01$$
$$p \approx 0.88$$

and $p[NBF(10)] + (1-p)[NBF(11)]$ will always be less than 0.01 since both $NBF(10)$ and $NBF(11)$ are less than 0.01. So, if the firm commits to using a threshold of 10 about 88% of the time and a threshold of 11 the rest of the time, truthful reporting will be incentive-compatible for the employees.

Under the original single threshold of 11, the firm never could fire a manager with exactly 11 misconduct interactions (since the single-threshold mechanism fires when there are more reports than the threshold). However, in the firm's ideal world, it always would fire a managers with exactly 11 misconduct interactions. Now, this

mixed mechanism allows the firm to fire a manager with 11 misconduct interactions 88% of the time, strictly improving upon the single-threshold mechanism.

In general, this improvement stems from the fact that the reports and thresholds need to be integers. In general, a reporting cost c_r will not fall precisely at the upper end of the range of allowed costs supporting truthful reporting under a given threshold. Mixing allows the firm to lower the expected threshold while maintaining incentive compatibility.

Example 4. *Mixing two thresholds can permit truthful reporting for reporting costs where no single threshold supports truthful reporting. A firm can also mix using a threshold that never supports truthful reporting in a single-threshold mechanism.*

The mixing of two thresholds creates a new range of reporting costs that support truthful reporting. Consider the parameters used before in Figure 3.2, where $\alpha = 5$, $\beta = 2$, $n = 20$, and $\bar{\gamma} = 0.5$. Now, in Figure 3.8, we can see that there is no truthful single-threshold mechanism for the reporting cost 0.01: it is above the allowed ranges for $m = 8$ and 9 and below the allowed ranges for $m \in [13, 19]$. Recall that for these parameters, the thresholds $m \in [10, 12]$ cannot support truthful reporting because $NBF > NBT$.

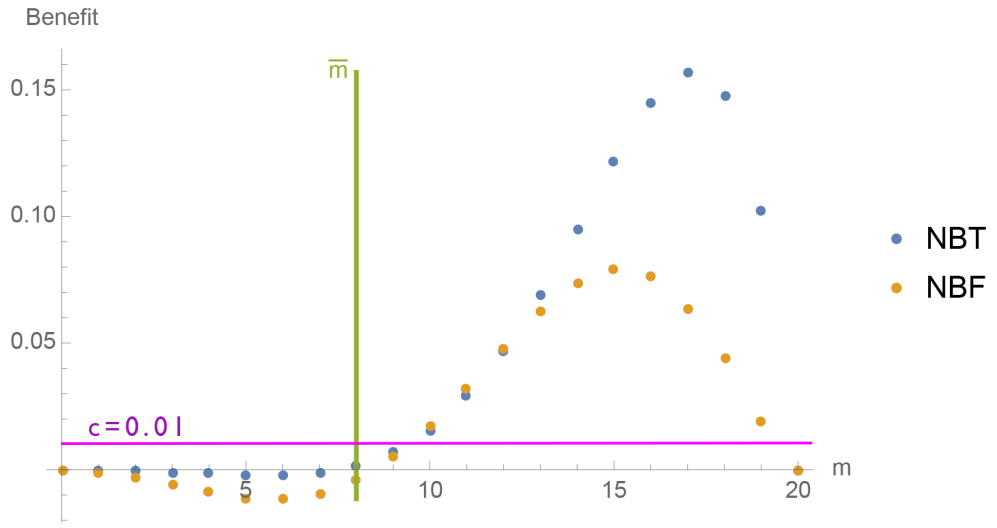


Figure 3.8: $\alpha = 5$, $\beta = 2$, $n = 20$, $\bar{\gamma} = 0.5$; reporting cost of 0.01 marked by the horizontal line.

Even though there are no thresholds such that $c_r = 0.01$ supports truthful reporting and there are no costs such that a threshold of 10 supports truthful reporting, mixing

between thresholds 9 and 10 can support truthful reporting for $c_r = 0.01$ To make those who experienced misconduct indifferent between reporting it and not, the firm can commit to using a threshold of 9 with a probability of approximately 65%. Mixing in those proportions leads to a net benefit of falsely reporting misconduct of 0.009, so truthful reporting is incentive-compatible. Remember that for these parameters, the firm would prefer any single-threshold mechanism with a threshold of 10 or below to solely relying on its prior belief about the manager and not eliciting information (though a single threshold of 10 is not possible with truthful reporting). So, this mixture of thresholds 9 and 10 allows the firm to create a mechanism for $c_r = 0.01$ that also is preferred to not eliciting information.

3.6 Discussion and Conclusion

We present a model of a firm seeking to minimize a manager's misconduct through an information escrow where the firm commits to fire the manager if the number of employees reporting misconduct surpasses a threshold. We aim to characterize when the firm can design such a mechanism so that truthful reporting is incentive compatible for the employees. Finally, we consider the firm's preference between different potential thresholds.

When reporting is costless and the firm and employees both seek to minimize misconduct, a single-threshold mechanism can reach firm's first-best in manner parallel to the setting of an aggregation rule for juror votes in Austen-Smith & Banks (1996). The optimal threshold will be the same as the firm's threshold if it could observe all of the manager's interactions with employees, and the employees will report their interactions truthfully.

However, when the firm and employee preferences diverge because it is costly for employees to report misconduct, the firm may no longer be able to achieve its first-best result. Truthful reporting depends on the pair of reporting cost and mechanism threshold. When the reporting cost is raised enough, the threshold for firing must be raised because an employee prefers to avoid paying the reporting cost by keeping some managers who are worse than the replacement hiring pool, and for high enough reporting costs, no employee will be willing to report misconduct and no truthful mechanism will exist. Less intuitively, when the prior belief about the manager is worse than the pool, it is possible for no truthful mechanism to exist for intermediate costs, where truthful mechanisms are possible for reporting costs both below and *above* those intermediate costs. This gap in costs — and a parallel gap in the

thresholds that support truthful reporting — stems from the net benefit of truthfully reporting misconduct falling below the net benefit of falsely reporting it. Employees who experience misconduct under a manager with a bad prior essentially have a free-riding incentive because they believe it is more likely that enough others also will have experienced misconduct to trigger the threshold.

Considering some design choices by the firm, we find that, for a given reporting cost, the firm prefers the truthful single-threshold mechanism with the lowest threshold. However, even the best single-threshold mechanism can be worse than 1) choosing not to elicit any information from the employees and 2) mixing two thresholds. A single-threshold mechanism, if it exists, will always improve on not eliciting information if the prior belief is that the manager is better than the pool. If the prior belief is that the manager is worse than the pool, the best single-threshold mechanism may be worse than firing the manager based on the prior belief and eliciting no information, because the mechanism threshold may be high enough that the firm loses more from retaining bad managers than it gains from keeping some good managers. Mixing over two thresholds generally dominates a single-threshold mechanism because of the discrete nature of the threshold. In addition, mixing over two thresholds can yield a truthful mechanism for reporting costs that have no truthful single-threshold mechanism.

The relationship between reporting costs and threshold may have policy implications for the design of information escrow systems. Measures to reduce the cost of reporting may lead to willingness to report under lower thresholds, closer to the ideal, as long as the range of reporting costs supporting truthful reporting is continuous. Information escrow systems can alter the cost of reporting by changing the requirements for making a report. For example, Callisto allows people to create time-stamped incident logs using a form based on the forensic experiential trauma interview, which although designed to minimize the amount that victims are re-traumatized is a fairly extensive and detailed interview. However, Callisto does not require people to fill out an incident log to enter its escrow system; it only requires a much less costly report including the state where the incident occurred, some information to identify the perpetrator, and contact information for the reporter.

The model offers several potential avenues for future research. We have shown several conditions under which single-threshold mechanisms are non-optimal, but further work remains to find the optimal threshold-based mechanism under costly reporting in this model. We also have focused on comparisons between different

escrow systems and between an escrow system and a decision based on prior beliefs about the manager, but we have not explored the comparison to individual employees deciding whether to report misconduct without the firm's commitment to an escrow with a firing threshold rule.

In addition, several extensions to the model would explore how the implications of costly reporting seen in this paper interact with other factors important to decisions about misconduct reporting. Firms face hiring costs for replacing managers. Firms also hire managers on the basis of managerial talent, not just their likelihood of committing misconduct. Extending the model to a two-dimensional manager, possessing both a talent level and a probability of misconduct, would help assess the impact of the firm's main hiring incentives. Employees may pay an additional retaliation-based cost if they report misconduct, but their report is not followed by the firing of the manager. Employees also could have heterogenous reporting costs and heterogeneous levels of animus against the manager, where they derive a benefit (or cost) from the firing of the manager unrelated to the manager's tendency to commit misconduct. While false reporting of misconduct can arise in our version of the model (i.e., when $NBF > c_r$), extending the model to include these factors may offer a more thorough picture of false reporting and the ability of an escrow to make truthful reporting incentive compatible.

BIBLIOGRAPHY

Abramowicz, M. (2001). A compromise approach to compromise verdicts. *California Law Review*, *89*(2), 231–314.

Alpern, S., & Gal, S. (2009). Analysis and design of selection committees: a game theoretic secretary problem. *International Journal of Game Theory*, *38*(3), 377–394.

Alpern, S., Gal, S., & Solan, E. (2010). A sequential selection game with vetoes. *Games and Economic Behavior*, *68*(1), 1–14.

American Bar Association. (1994). *Aba standards for criminal justice: Sentencing*. Retrieved from `https://www.americanbar.org/content/dam/aba/publications/criminal_justice_standards/sentencing.pdf` (accessed May 8, 2021)

Andreoni, J. (1991). Reasonable doubt and the optimal magnitude of fines: Should the penalty fit the crime? *The RAND Journal of Economics*, *22*(3), 385–395.

Anwar, S., Bayer, P., & Hjalmarsson, R. (2012). The impact of jury race in criminal trials. *The Quarterly Journal of Economics*, *127*(2), 1017–1055.

APSA Council. (2018). Minutes of march 2018, apsa council meeting. *PS: Political Science & Politics*, *51*(4), 920–924.

Arun, V., Kate, A., Garg, D., Druschel, P., & Bhattacharjee, B. (2018). Finding safety in numbers with secure allegation escrows. *arXiv preprint arXiv:1810.10123*.

Austen-Smith, D., & Banks, J. S. (1996). Information aggregation, rationality, and the Condorcet jury theorem. *American political science review*, *90*(1), 34–45.

Ayres, I. (2017). Voluntary taxation and beyond: The promise of social-contracting voting mechanisms. *American Law and Economics Review*, *19*(1), 1–48.

Ayres, I. (2018). Targeting repeat offender ndas. *Stanford Law Review Online*, *71*, 76.

Ayres, I., Chwe, M., & Ladd, J. (2017). Act-sampling bias and the shrouding of repeat offending. *Virginia Law Review Online*, *103*, 94.

Ayres, I., & Unkovic, C. (2012). Information escrows. *Michigan Law Review*, *111*, 145–196.

Babcock, L., & Landeo, C. M. (2004). Settlement escrows: An experimental study of a bilateral bargaining game. *Journal of Economic Behavior & Organization*, *53*(3), 401–417.

Barak-Corren, N., & Lewinsohn-Zamir, D. (2019). What's in a name? the disparate effects of identifiability on offenders and victims of sexual harassment. *Journal of Empirical Legal Studies*, *16*(4), 955–1000.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, *76*(2), 169–217.

Bellin, J. (2010). Is punishment relevant after all? a prescription for informing juries of the consequences of conviction. *Boston University Law Review*, *90*, 2223–2265.

Bindler, A., & Hjalmarsson, R. (2018). How punishment severity affects jury verdicts: Evidence from two natural experiments. *American Economic Journal: Economic Policy*, *10*(4), 36–78.

Blanck, P. D. (1993). Calibrating the scales of justice: Studying judges' behaviors in bench trials. *Indiana Law Journal*, *68*(4), 1119–1198.

Boudreau, L. E., Chassang, S., Gonzalez-Torres, A., Heath, R., & of Economic Research, N. B. (2023). *Monitoring harassment in organizations* (Tech. Rep.). National Bureau of Economic Research.

Brams, S. J., & Davis, M. D. (1978). Optimal jury selection: A game-theoretic model for the exercise of peremptory challenges. *Operations research*, *26*(6), 966–991.

Bushway, S. D., Owens, E. G., & Piehl, A. M. (2012). Sentencing guidelines and judicial discretion: Quasi-experimental evidence from human calculation errors. *Journal of Empirical Legal Studies*, *9*(2), 291–319.

Butler, P. (1995). Racially based jury nullification: Black power in the criminal justice system. *The Yale Law Journal*, *105*, 677–725.

Cantalupo, N. C., & Kidder, W. C. (2018). A systematic look at a serial problem: Sexual harassment of students by university faculty. *Utah Law Review*, 671.

Carrington, M. (2011). Note, applying *Apprendi* to jury sentencing: Why state felony jury sentencing threatens the right to a jury trial. *U. Ill. L. Rev.*, *2011*, 1359–1385.

Carrubba, C. J., & Clark, T. S. (2012). Rule creation in a political hierarchy. *The American Political Science Review*, *106*(3), 622–643.

Chassang, S., & Miquel, G. P. I. (2019). Crime, intimidation, and whistleblowing: A theory of inference from unverifiable reports. *The Review of Economic Studies*, *86*(6), 2530–2553.

Chassang, S., & Zehnder, C. (2019). *Secure survey design in organizations: Theory and experiments* (Tech. Rep.). National Bureau of Economic Research. (Working Paper 25918)

Cheng, I.-H., & Hsiaw, A. (2022). Reporting sexual misconduct in the # metoo era. *American Economic Journal: Microeconomics*, *14*(4), 761–803.

Cohen, A., Klement, A., & Neeman, Z. (2015). Judicial decision making: A dynamic reputation approach. *The Journal of Legal Studies*, *44*(S1), 133–159.

Craft, W. (2018). Peremptory strikes in Mississippi's Fifth Circuit Court District. *APM Reports*. Retrieved from `https://features.apmreports.org/files/peremptory_strike_methodology.pdf` (data available at `https://github.com/APM-Reports/jury-data/` and licensed under CC BY 4.0, `https://creativecommons.org/licenses/by/4.0/legalcode`)

De Clippel, G., Eliaz, K., & Knight, B. (2014). On the selection of arbitrators. *American Economic Review*, *104*(11), 3434–58.

DeGroot, M. H., & Kadane, J. B. (1980). Optimal challenges for selection. *Operations Research*, *28*(4), 952–968.

Diamond, S. S. (2007). Dispensing with deception, curing with care: A response to judge dann on nullification. *Judicature*, *91*(1), 20–25.

Dietrich, B. J., Enos, R. D., & Sen, M. (2019). Emotional arousal predicts voting on the u.s. supreme court. *Political Analysis*, *27*, 237–243.

Dobbin, F., & Kalev, A. (2020). Making discrimination and harassment complaint systems better. In C. for Employment Equity (Ed.), *What works?: Evidence-based ideas to increase diversity, equity, and inclusion in the workplace* (pp. 24–29). University of Massachusetts Amherst. Retrieved from `https://www.umass.edu/employmentequity/what-works`

Duggan, J., & Martinelli, C. (2001). A Bayesian model of voting in juries. *Games and Economic Behavior*, *37*(2), 259–294.

Duvall, K. (2012). The contradictory stance on jury nullification. *North Dakota Law Review*, *88*(2), 409–452.

Eisenberg, T., Hannaford-Agor, P. L., Hans, V. P., Waters, N. L., Munsterman, G. T., Schwab, S. J., & Wells, M. T. (2005). Judge-jury agreement in criminal cases: A partial replication of kalven and zeisel's *The American Jury*. *Journal of Empirical Legal Studies*, *2*(1), 171–206.

Epps, D., & Ortman, W. (2022). The informed jury. *Vanderbilt Law Review*, *75*(3), 823–890.

Feddersen, T., & Pesendorfer, W. (1998). Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting. *American Political science review*, *92*(1), 23–35.

Fisher, T. (2011). Constitutionalism and the criminal law: Rethinking criminal trial bifurcation. *The University of Toronto Law Journal*, *61*(4), 811–843.

Flanagan, F. X. (2015). Peremptory challenges and jury selection. *The Journal of Law and Economics*, *58*(2), 385–416.

Flanagan, F. X. (2018). Race, gender, and juries: Evidence from North Carolina. *The Journal of Law and Economics*, *61*(2), 189–214.

Ford, R. A. (2009). Modeling the effects of peremptory challenges on jury selection and jury verdicts. *Geo. Mason L. Rev.*, *17*, 377.

Garvey, S. P., Hannaford-Agor, P., Hans, V. P., Mott, N. L., Munsterman, G. T., & Wells, M. T. (2004). Juror first votes in criminal trials. *Journal of Empirical Legal Studies*, *1*(2), 371–398.

Gau, J. M. (2016). A jury of whose peers? the impact of selection procedures on racial composition and the prevalence of majority-white juries. *Journal of Crime and Justice*, *39*(1), 75–87.

Gay, G. D., Grace, M. F., Kale, J. R., & Noe, T. H. (1989). Noisy juries and the choice of trial mode in a sequential signalling game: Theory and evidence. *The RAND Journal of Economics*, *20*(2), 196–213.

Gertner, R. H., & Miller, G. P. (1995). Settlement escrows. *The Journal of Legal Studies*, *24*(1), 87–122.

Grover, M. (2019). Note, jury sentencing in the united states: The antithesis of the rule of law. *Mitchell Hamline L.J. Pub. Pol'y & Prac.*, *40*, 23–50.

Guttel, E., & Teichman, D. (2012). Criminal sanctions in the defense of the innocent. *Michigan Law Review*, *110*(4), 597–645.

Hagerty, B. B. (2019, August). An epidemic of disbelief. *The Atlantic*.

Hannaford-Agor, P., & Hans, V. P. (2003). Nullification at work? a glimpse from the national center for state courts study of hung juries. *Chicago-Kent Law Review*, *78*, 1249–1277.

Hemel, D., & Lund, D. S. (2018). Sexual harassment and corporate law. *Columbia Law Review*, *118*(6), 1583–1680.

Hoffman, M. (2003). The case for jury sentencing. *Duke L.J.*, *52*, 951–1010.

Holtzman, K. A. (2021). Note, criminal advisory juries: A sensible compromise for jury sentencing advocates. *N.W. J.L. & Soc. Pol'y*, *16*, 164–191.

Horowitz, I. A. (2008). Jury nullification: An empirical perspective. *Northern Illinois University Law Review*, *28*(3), 425–452.

Hurston, N. (2023, December 26). High praise, hard questions: Legislators interview new, returning judicial candidates. *Virginia Lawyers Weekly*.

Iontcheva, J. (2003). Jury sentencing as democratic practice. *Va. L. Rev.*, *89*, 311–383.

Kadane, J. B., Stone, C. A., & Wallstrom, G. (1999). The donation paradox for peremptory challenges. *Theory and Decision*, *47*(2), 139–155.

Kalven, H., & Zeisel, H. (1966). *The american jury*. Little, Brown.

Kaplan, M. F., & Krupa, S. (1986). Severe penalties under the control of others can reduce guilt verdicts. *Law and Psychology Review*, *10*, 1–18.

King, N. J. (2004). How different is death? jury sentencing in capital and non-capital cases compared. *Ohio State Journal of Criminal Law*, *2*, 195–214.

King, N. J., & Noble, R. L. (2005). Jury sentencing in noncapital cases: Comparing severity and variance with judicial sentences in two states. *Journal of Empirical Legal Studies*, *2*(2), 331–367.

Klein, A. L. (2021). Meaningless guarantees: Comment on mitchell e. mccloy's "blind justice: Virginia's jury sentencing scheme and impermissible burdens on a defendant's right to a jury trial". *Washington and Lee Law Review*, *78*(1), 585–598.

Kling, J. R. (2006). Incarceration length, employment, and earnings. *The American Economic Review*, *96*(3), 863–876.

Lanni, A. (1999). Jury sentencing in noncapital cases: An idea whose time has come (again)? *Yale L.J.*, *82*, 1775–1803.

Lee, F. X., & Suen, W. (2020). Credibility of crime allegations. *American Economic Journal: Microeconomics*, *12*(1), 220–259.

Lehmann, J.-Y. K., & Smith, J. B. (2013). *Power to bias? the effect of attorney empowerment in voir dire on jury prejudice and race.* (Working Paper)

Leipold, A. D. (2005). Why are federal judges so acquittal prone? *Washington University Law Quarterly*, *83*, 151–227.

Lempert, R. O. (2001). The economic analysis of evidence law: Common sense on stilts. *Virginia Law Review*, *87*, 1619–1712.

Leshem, E. A. (2019). Jury selection as election: A new framework for peremptory strikes. *Yale L. J.*, *128*, 2356.

Liptak, A. (2015, August 16). Exclusion of blacks from juries raises renewed scrutiny. *New York Times*.

Lovell, R., Luminais, M., Flannery, D. J., Bell, R., & Kyker, B. (2018). Describing the process and quantifying the outcomes of the cuyahoga county sexual assault kit initiative. *Journal of Criminal Justice*, *57*, 106–115.

Lundberg, A. (2016a). *Do judges vary their burden of proof? evidence from federal bench trials.* (Working Paper)

Lundberg, A. (2016b). Sentencing discretion and burdens of proof. *International Review of Law and Economics*, *46*, 34–42.

McCloy, M. E. (2021). Note, blind justice: Virginia's jury sentencing scheme and impermissible burdens on a defendant's right to a jury trial. *Wash. & Lee L. Rev.*, *78*, 519–584.

Miceli, T. J. (2008). Criminal sentencing guidelines and judicial discretion. *Contemporary Economic Policy*, *26*(2), 207–215.

Moro, A., & Van der Linden, M. (2022). *Exclusion of extreme jurors and minority representation: The effect of jury selection procedures.* (Working Paper)

Munsterman, G. T., Hannaford-Agor, P. L., & Whitehead, G. M. (Eds.). (2006). *Jury trial innovations* (2nd ed.). National Center for State Courts.

Neilson, W. S., & Winter, H. (2000). Bias and the economics of jury selection. *International Review of Law and Economics*, *20*(2), 223–250.

Parameswaran, G., Cameron, C. M., & Kornhauser, L. A. (2021). Bargaining and strategic voting on appellate courts. *The American Political Science Review*, 1–16.

Pei, H., & Strulovici, B. (2021). *Crime aggregation, deterrence, and witness credibility.* (Working Paper)

Pew Research Center. (2021, October 26). *Growing share of americans say they want more spending on police in their area.* Retrieved from https://www.pewresearch.org/short-reads/2021/10/26/growing-share-of-americans-say-they-want-more-spending-on-police-in-their-area/ (accessed August 20, 2023)

Polinsky, A. M., & Shavell, S. (2000). The fairness of sanctions: Some implications for optimal enforcement policy. *American Law and Economics Review*, 223–237.

Posner, R. A. (1999). An economic approach to the law of evidence. *Stanford Law Review*, *51*(6), 1477–1546.

President's Commission on Law Enforcement and Administration of Justice. (1967). *The challenge of crime in a free society: A report by the president's commission on law enforcement and administration of justice.* Retrieved from https://www.ojp.gov/sites/g/files/xyckuh241/files/archives/ncjrs/42.pdf (accessed May 8, 2021)

Rajan, A., Qin, L., Archer, D. W., Boneh, D., Lepoint, T., & Varia, M. (2018). Callisto: A cryptographic approach to detecting serial perpetrators of sexual misconduct. In *Proceedings of the 1st acm sigcas conference on computing and sustainable societies* (pp. 1–4).

Rappaport, J. (2020). Some doubts about "democratizing" criminal justice. *The University of Chicago Law Review*, *87*(3), 711–813.

Reinganum, J. F. (2000). Sentencing guidelines, judicial discretion, and plea bargaining. *The RAND Journal of Economics*, *31*(1), 62–81.

Revesz, J. (2016). Ideological imbalance and the peremptory challenge. *Yale L.J.*, *125*, 2535. (Comment)

Rose, M. R. (1999). The peremptory challenge accused of race or gender discrimination? some data from one county. *Law and Human Behavior*, *23*(6), 695.

Roth, A., Kadane, J. B., & Degroot, M. H. (1977). Optimal peremptory challenges in trials by juries: a bilateral sequential process. *Operations Research*, *25*(6), 901–919.

Salib, P. N., & Krishnamurthi, G. (2022). Jury nullification in abortion prosecutions: An equilibrium theory. *Duke Law Journal Online*, *72*, 41–57.

Scapicchio, R. (2022, November 7). Peremptory challenge should be reserved for the defendant. *Boston Bar Journal*, *66*(4). Retrieved from `https://bostonbar.org/journal/peremptorychallenge`

Schuman, J. (2015). Probability and punishment: How to improve sentencing by taking account of probability. *New Criminal Law Review*, *18*(2), 214–272.

Schwartz, E. P., & Schwartz, W. F. (1996). The challenge of peremptory challenges. *The Journal of Law, Economics, and Organization*, *12*(2), 325–360.

Shadmehr, M., Cameron, C., & Shahshahani, S. (2022). Coordination and innovation in judiciaries: Correct law vs. consistent law. *Quarterly Journal of Political Science*, *17*(1), 61–89.

Shavell, S. (2007). Optimal discretion in the application of rules. *American Law and Economics Review*, *9*(1), 175–194.

Shepherd, J. M. (2002). Police, prosecutors, criminals, and determinate sentencing: The truth about truth-in-sentencing laws. *Journal of Law and Economics*, *45*, 509–534.

Siegel, R., & Strulovici, B. (2019). *The economic case for probability-based sentencing.* (Working Paper)

Silveira, B. S. (2017). Bargaining with asymmetric information: An empirical study of plea negotiations. *Econometrica*, *85*(2), 419–452.

Teichman, D. (2017). Convicting with reasonable doubt: An evidentiary theory of criminal law. *Notre Dame Law Review*, *93*(2), 757–810.

A.B. 3070. (Ca. 2020). Retrieved from https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB3070

Apprendi v. New Jersey. (2000). 530 U.S. 466.

Ariz. R. Civ. P. 47. (n.d.).

Ariz. R. Crim. P. 18. (n.d.).

Batson v. Kentucky. (1986). 476 U.S. 79.

Miss. R. Crim. P. 18. (n.d.).

People v. Kriho. (1999). 996 P.2d 158.

Ring v. Arizona. (2002). 536 U.S. 584.

Wash. Gen. R. 37. (n.d.).

Trial Partners, I. (2012). *What every practitioner should know about jury selection and* Voir Dire *in employment cases.* Retrieved from http://www.trial-partners.com/wp-content/uploads/2017/07/Voir-Dire-Cases.pdf (accessed June 13, 2020)

Tuerkheimer, D. (2019). Beyond # metoo. *NYU Law Review*, *94*, 1146.

Van der Linden, M. (2017). *Strategic simplicity in jury selection, committee selection, and matching* (Unpublished doctoral dissertation). Vanderbilt University.

Webster, C. W. (1960). Jury sentencing — grab-bag justice. *Sw. L.J.*, *14*, 221–30.

Weinstein, J. B. (1992). A trial judge's second impression of the federal sentencing guidelines. *Southern California Law Review*, *66*(1), 357–366.

Weninger, R. A. (1994). Jury sentencing in noncapital cases: A case study of el paso county, texas. *Journal of Urban and Contemporary Law*, *45*(3), 3–40.

Widman, L., Olson, M. A., & Bolen, R. M. (2013). Self-reported sexual assault in convicted sex offenders and community men. *Journal of Interpersonal Violence*, *28*(7), 1519–1536.

Appendix A

MORE RESULTS FOR GENERAL CASE ATTORNEY BEHAVIOR

Jurors 2 and 3 from the Same Group

Consider when $g_2 = g_3$, and relabel the two possible values of π_2, π_3 as $\underline{g}, \overline{g}$. Then, half of the time, when π_2 is realized it will equal $\underline{g}$ and

$$\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{g}] = \begin{cases} \left(\frac{\nu}{\gamma}\frac{\underline{g}}{1-\underline{g}}\frac{1-\rho}{\rho}\right)^{\frac{\nu}{\gamma-\nu}} & \text{if } \omega = 0 \\ \left(\frac{\nu}{\gamma}\frac{\underline{g}}{1-\underline{g}}\frac{1-\rho}{\rho}\right)^{\frac{\gamma}{\gamma-\nu}} & \text{if } \omega = 1 \end{cases}$$

The other half of the time, $\pi_2 = \overline{g}$ and

$$\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \overline{g}] = \begin{cases} \left(\frac{\nu}{\gamma}\frac{\overline{g}}{1-\overline{g}}\frac{1-\rho}{\rho}\right)^{\frac{\nu}{\gamma-\nu}} & \text{if } \omega = 0 \\ \left(\frac{\nu}{\gamma}\frac{\overline{g}}{1-\overline{g}}\frac{1-\rho}{\rho}\right)^{\frac{\gamma}{\gamma-\nu}} & \text{if } \omega = 1 \end{cases}$$

The attorney making the choice to strike juror 2 will compare those realized values with the expected payoff given g_3:

$$\mathbb{P}[s_3 > s_3^* | \omega, g_3] = \begin{cases} \frac{1}{2}\left(\frac{\nu}{\gamma}\frac{\underline{g}}{1-\underline{g}}\frac{1-\rho}{\rho}\right)^{\frac{\nu}{\gamma-\nu}} + \frac{1}{2}\left(\frac{\nu}{\gamma}\frac{\overline{g}}{1-\overline{g}}\frac{1-\rho}{\rho}\right)^{\frac{\nu}{\gamma-\nu}} & \text{if } \omega = 0 \\ \frac{1}{2}\left(\frac{\nu}{\gamma}\frac{\underline{g}}{1-\underline{g}}\frac{1-\rho}{\rho}\right)^{\frac{\gamma}{\gamma-\nu}} + \frac{1}{2}\left(\frac{\nu}{\gamma}\frac{\overline{g}}{1-\overline{g}}\frac{1-\rho}{\rho}\right)^{\frac{\gamma}{\gamma-\nu}} & \text{if } \omega = 1 \end{cases}$$

Note that since $\gamma - \nu < 0$ and $\underline{g} \leq \overline{g}$,

$$\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{g}] \geq \mathbb{P}[s_3 > s_3^* | \omega, g_3] \geq \mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \overline{g}]$$

So,

$$\begin{aligned} &\mathbb{E}[\max(\mathbb{P}[s_2 > s_2^* | \omega, \pi_2], \mathbb{P}[s_3 > s_3^* | \omega, g_3])] \\ &\qquad = \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{g}] + \frac{1}{2}\mathbb{P}[s_3 > s_3^* | \omega, g_3] \\ &\qquad = \begin{cases} \frac{3}{4}\left(\frac{\nu}{\gamma}\frac{\underline{g}}{1-\underline{g}}\frac{1-\rho}{\rho}\right)^{\frac{\nu}{\gamma-\nu}} + \frac{1}{4}\left(\frac{\nu}{\gamma}\frac{\overline{g}}{1-\overline{g}}\frac{1-\rho}{\rho}\right)^{\frac{\nu}{\gamma-\nu}} & \text{if } \omega = 0 \\ \frac{3}{4}\left(\frac{\nu}{\gamma}\frac{\underline{g}}{1-\underline{g}}\frac{1-\rho}{\rho}\right)^{\frac{\gamma}{\gamma-\nu}} + \frac{1}{4}\left(\frac{\nu}{\gamma}\frac{\overline{g}}{1-\overline{g}}\frac{1-\rho}{\rho}\right)^{\frac{\gamma}{\gamma-\nu}} & \text{if } \omega = 1 \end{cases} \end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[\min(\mathbb{P}[s_2 > s_2^*|\omega,\pi_2],\mathbb{P}[s_3 > s_3^*|\omega,g_3])] \\ &= \frac{1}{2}\mathbb{P}[s_2 > s_2^*|\omega,\pi_2=\overline{g}] + \frac{1}{2}\mathbb{P}[s_3 > s_3^*|\omega,g_3] \\ &= \begin{cases} \frac{1}{4}\left(\frac{\nu}{\gamma}\frac{\underline{g}}{1-\underline{g}}\frac{1-\rho}{\rho}\right)^{\frac{\nu}{\gamma-\nu}} + \frac{3}{4}\left(\frac{\nu}{\gamma}\frac{\overline{g}}{1-\overline{g}}\frac{1-\rho}{\rho}\right)^{\frac{\nu}{\gamma-\nu}} & \text{if } \omega = 0 \\ \frac{1}{4}\left(\frac{\nu}{\gamma}\frac{\underline{g}}{1-\underline{g}}\frac{1-\rho}{\rho}\right)^{\frac{\gamma}{\gamma-\nu}} + \frac{3}{4}\left(\frac{\nu}{\gamma}\frac{\overline{g}}{1-\overline{g}}\frac{1-\rho}{\rho}\right)^{\frac{\gamma}{\gamma-\nu}} & \text{if } \omega = 1 \end{cases}\end{aligned}$$

Jurors 2 and 3 from Different Groups

Consider when $g_2 = A$ and $g_3 = B$. Then, half of the time, when π_2 is realized it will equal $\underline{a}$ and

$$\mathbb{P}[s_2 > s_2^*|\omega,\pi_2=\underline{a}] = \begin{cases} \left(\frac{\nu}{\gamma}\frac{\underline{a}}{1-\underline{a}}\frac{1-\rho}{\rho}\right)^{\frac{\nu}{\gamma-\nu}} & \text{if } \omega = 0 \\ \left(\frac{\nu}{\gamma}\frac{\underline{a}}{1-\underline{a}}\frac{1-\rho}{\rho}\right)^{\frac{\gamma}{\gamma-\nu}} & \text{if } \omega = 1 \end{cases}$$

The other half of the time, $\pi_2 = \overline{a}$ and

$$\mathbb{P}[s_2 > s_2^*|\omega,\pi_2=\overline{a}] = \begin{cases} \left(\frac{\nu}{\gamma}\frac{\overline{a}}{1-\overline{a}}\frac{1-\rho}{\rho}\right)^{\frac{\nu}{\gamma-\nu}} & \text{if } \omega = 0 \\ \left(\frac{\nu}{\gamma}\frac{\overline{a}}{1-\overline{a}}\frac{1-\rho}{\rho}\right)^{\frac{\gamma}{\gamma-\nu}} & \text{if } \omega = 1 \end{cases}$$

The attorney making the choice to strike juror 2 will compare those realized values with the expected payoff given g_3:

$$\mathbb{P}[s_3 > s_3^*|\omega,g_3=B] = \begin{cases} \frac{1}{2}\left(\frac{\nu}{\gamma}\frac{\underline{b}}{1-\underline{b}}\frac{1-\rho}{\rho}\right)^{\frac{\nu}{\gamma-\nu}} + \frac{1}{2}\left(\frac{\nu}{\gamma}\frac{\overline{b}}{1-\overline{b}}\frac{1-\rho}{\rho}\right)^{\frac{\nu}{\gamma-\nu}} & \text{if } \omega = 0 \\ \frac{1}{2}\left(\frac{\nu}{\gamma}\frac{\underline{b}}{1-\underline{b}}\frac{1-\rho}{\rho}\right)^{\frac{\gamma}{\gamma-\nu}} + \frac{1}{2}\left(\frac{\nu}{\gamma}\frac{\overline{b}}{1-\overline{b}}\frac{1-\rho}{\rho}\right)^{\frac{\gamma}{\gamma-\nu}} & \text{if } \omega = 1 \end{cases}$$

Here, again, we always have that

$$\mathbb{P}[s_2 > s_2^*|\omega,\pi_2=\underline{a}] \geq \mathbb{P}[s_2 > s_2^*|\omega,\pi_2=\overline{a}]$$

If $\mathbb{P}[s_3 > s_3^*|\omega,g_3=B] > \mathbb{P}[s_2 > s_2^*|\omega,\pi_2=\underline{a}] \geq \mathbb{P}[s_2 > s_2^*|\omega,\pi_2=\overline{a}]$, then

$$\begin{aligned}\mathbb{E}[\max(\mathbb{P}[s_2 > s_2^*|\omega,\pi_2],\mathbb{P}[s_3 > s_3^*|\omega,g_3])] \\ &= \mathbb{P}[s_3 > s_3^*|\omega,g_3=B] \\ \mathbb{E}[\min(\mathbb{P}[s_2 > s_2^*|\omega,\pi_2],\mathbb{P}[s_3 > s_3^*|\omega,g_3])] \\ &= \frac{1}{2}\mathbb{P}[s_2 > s_2^*|\omega,\pi_2=\underline{a}] + \frac{1}{2}\mathbb{P}[s_2 > s_2^*|\omega,\pi_2=\overline{a}]\end{aligned}$$

If $\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{a}] \geq \mathbb{P}[s_3 > s_3^* | \omega, g_3 = B] > \mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \overline{a}]$, then

$$\mathbb{E}[\max(\mathbb{P}[s_2 > s_2^* | \omega, \pi_2], \mathbb{P}[s_3 > s_3^* | \omega, g_3])] = \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{a}] + \frac{1}{2}\mathbb{P}[s_3 > s_3^* | \omega, g_3 = B]$$

$$\mathbb{E}[\min(\mathbb{P}[s_2 > s_2^* | \omega, \pi_2], \mathbb{P}[s_3 > s_3^* | \omega, g_3])] = \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \overline{a}] + \frac{1}{2}\mathbb{P}[s_3 > s_3^* | \omega, g_3 = B]$$

If $\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{a}] \geq \mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \overline{a}] \geq \mathbb{P}[s_3 > s_3^* | \omega, g_3 = B]$, then

$$\mathbb{E}[\max(\mathbb{P}[s_2 > s_2^* | \omega, \pi_2], \mathbb{P}[s_3 > s_3^* | \omega, g_3])] = \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{a}] + \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \overline{a}]$$

$$\mathbb{E}[\min(\mathbb{P}[s_2 > s_2^* | \omega, \pi_2], \mathbb{P}[s_3 > s_3^* | \omega, g_3])] = \mathbb{P}[s_3 > s_3^* | \omega, g_3 = B]$$

Note that it is possible to be in one of these cases when $\omega = 0$ and a different case when $\omega = 1$. In other words, the maximum and minimum values may be different when the defendant is guilty versus when the defendant is not.

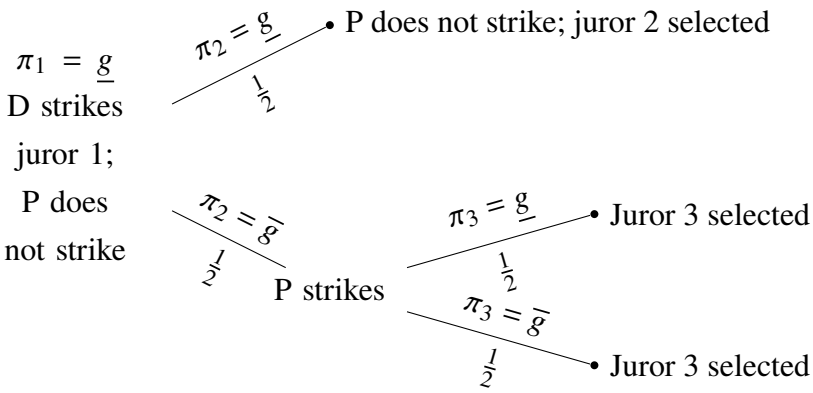
Appendix B

GENERAL CASE JUROR DISTRIBUTIONS

Jurors 2 and 3 from the Same Group

Again, let $g_2 = g_3$, and relabel the two possible values of π_2, π_3 as $\underline{g}, \overline{g}$. Label the other group's payoff values $\underline{G}, \overline{G}$. We consider cases based on the realization of the first juror's payoff: $\pi_1 \in \{\underline{g}, \overline{g}, \underline{G}, \overline{G}\}$.

If $\pi_1 = \underline{g}$, then



Thus, when $\pi_1 = \underline{g}$, the probability that the selected juror has each of the possible payoff values is as follows:

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
$\frac{3}{4}$	$\frac{1}{4}$	0	0

When $\pi_1 = \overline{g}$,

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
$\frac{1}{4}$	$\frac{3}{4}$	0	0

When $\pi_1 = G \in \{\underline{G}, \overline{G}\}$, attorney behavior depends on the value of G. Let

$$x_\omega = \begin{cases} \nu & \text{if } \omega = 0 \\ \gamma & \text{if } \omega = 1 \end{cases}$$

Then, the defense attorney strikes juror 1 when the known payoff from juror 1 is worse than the expected payoff from letting the prosecutor choose between a known payoff of juror 2 and the expected payoff of juror 3

$$1-\left(\frac{\nu}{\gamma}\frac{G}{1-G}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}} < 1-\frac{3}{4}\left(\frac{\nu}{\gamma}\frac{\underline{g}}{1-\underline{g}}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}}-\frac{1}{4}\left(\frac{\nu}{\gamma}\frac{\overline{g}}{1-\overline{g}}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}}$$

$$\left(\frac{\nu}{\gamma}\frac{G}{1-G}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}} > \frac{3}{4}\left(\frac{\nu}{\gamma}\frac{\underline{g}}{1-\underline{g}}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}}+\frac{1}{4}\left(\frac{\nu}{\gamma}\frac{\overline{g}}{1-\overline{g}}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}}$$

and the probabilities for the selected juror's payoff values are the same as when the defense strikes for $\pi_1 = \underline{g}$

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
$\frac{3}{4}$	$\frac{1}{4}$	0	0

The prosecutor strikes juror 1 when

$$\left(\frac{\nu}{\gamma}\frac{G}{1-G}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}} < \frac{1}{4}\left(\frac{\nu}{\gamma}\frac{\underline{g}}{1-\underline{g}}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}}+\frac{3}{4}\left(\frac{\nu}{\gamma}\frac{\overline{g}}{1-\overline{g}}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}}$$

and thus we have

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
$\frac{1}{4}$	$\frac{3}{4}$	0	0

Neither strikes juror 1 when

$$\left(\frac{\nu}{\gamma}\frac{G}{1-G}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}} \in \left[\frac{1}{4}\left(\frac{\nu}{\gamma}\frac{\underline{g}}{1-\underline{g}}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}}+\frac{3}{4}\left(\frac{\nu}{\gamma}\frac{\overline{g}}{1-\overline{g}}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}},\right.$$
$$\left.\frac{3}{4}\left(\frac{\nu}{\gamma}\frac{\underline{g}}{1-\underline{g}}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}}+\frac{1}{4}\left(\frac{\nu}{\gamma}\frac{\overline{g}}{1-\overline{g}}\frac{1-\rho}{\rho}\right)^{\frac{x_\omega}{\gamma-\nu}}\right]$$

and we have

$\underline{g}$	$\overline{g}$	G	$\sim G$
0	0	1	0

Jurors 2 and 3 from Different Groups

Let g_2 and g_3 be different, and relabel the two possible values of π_2 as $\underline{g}, \overline{g}$ and those of g_3 as $\underline{G}, \overline{G}$. As noted in section A, the expected maximum and minimum of the realized probability of conviction for juror 2 and the expected probability of conviction for juror 3 will change based on the relation between $\mathbb{P}[s_3 > s_3^*|\omega, g_3 = G]$, $\mathbb{P}[s_2 > s_2^*|\omega, \pi_2 = \underline{g}]$, and $\mathbb{P}[s_2 > s_2^*|\omega, \pi_2 = \overline{g}]$.

Case 1 : $\mathbb{P}[s_3 > s_3^*|\omega, g_3 = G] > \mathbb{P}[s_2 > s_2^*|\omega, \pi_2 = \underline{g}] \geq \mathbb{P}[s_2 > s_2^*|\omega, \pi_2 = \overline{g}]$

In this case, when choosing between jurors 2 and 3, the defense attorney will always choose to keep juror 2 and the prosecutor will always choose to strike juror 2. So, the defense attorney will strike juror 1 if

$$1 - \mathbb{P}[s_1 > s_1^*|\omega, \pi_1] < 1 - \mathbb{P}[s_3 > s_3^*|\omega, g_3 = G]$$
$$1 - \mathbb{P}[s_1 > s_1^*|\omega, \pi_1] < 1 - \frac{1}{2}\mathbb{P}[s_3 > s_3^*|\omega, \pi_3 = \underline{G}] - \frac{1}{2}\mathbb{P}[s_3 > s_3^*|\omega, \pi_3 = \overline{G}]$$

which means that the defense attorney will not strike when $\pi_1 = \underline{g}, \overline{g}$, or $\overline{G}$ and will strike only when $\pi_1 = \underline{G}$.

The prosecutor will strike juror 1 if

$$\mathbb{P}[s_1 > s_1^*|\omega, \pi_1] < \frac{1}{2}\mathbb{P}[s_2 > s_2^*|\omega, \pi_2 = \underline{g}] + \frac{1}{2}\mathbb{P}[s_2 > s_2^*|\omega, \pi_2 = \overline{g}]$$

which means that the prosecutor will not strike when $\pi_1 = \underline{g}$ or $\underline{G}$ and will strike when $\pi_1 = \overline{g}$. If $\pi_1 = \overline{G}$, then the prosecutor will strike if $\mathbb{P}[s_1 > s_1^*|\omega, \pi_1 = \overline{G}] < \frac{1}{2}\mathbb{P}[s_2 > s_2^*|\omega, \pi_2 = \underline{g}] + \frac{1}{2}\mathbb{P}[s_2 > s_2^*|\omega, \pi_2 = \overline{g}]$ and otherwise will not.

So, we have the following distributions for different values of π_1: when $\pi_1 = \underline{g}$ (no strikes),

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
1	0	0	0

when $\pi_1 = \overline{g}$ (prosecutor strikes, defense keeps juror 2),

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
$\frac{1}{2}$	$\frac{1}{2}$	0	0

when $\pi_1 = \underline{G}$ (defense strikes, then prosecutor strikes),

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
0	0	$\frac{1}{2}$	$\frac{1}{2}$

when $\pi_1 = \overline{G}$ and $\mathbb{P}[s_1 > s_1^* | \omega, \pi_1 = \overline{G}] < \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{g}] + \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \overline{g}]$ (prosecutor strikes, defense keeps juror 2),

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
$\frac{1}{2}$	$\frac{1}{2}$	0	0

and when $\pi_1 = \overline{G}$ and $\mathbb{P}[s_1 > s_1^* | \omega, \pi_1 = \overline{G}] \geq \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{g}] + \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \overline{g}]$ (no strikes),

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
0	0	0	1

Case 2: $\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{g}] \geq \mathbb{P}[s_3 > s_3^* | \omega, g_3 = G] > \mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \overline{g}]$

In this case, when choosing between jurors 2 and 3, the defense attorney will strike juror 2 if $\pi_2 = \underline{g}$ and keep juror 2 if $\pi_2 = \overline{g}$. The prosecutor will do the reverse. So, the defense attorney will strike juror 1 if

$$1 - \mathbb{P}[s_1 > s_1^* | \omega, \pi_1] < 1 - \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{g}] - \frac{1}{2}\mathbb{P}[s_3 > s_3^* | \omega, g_3 = G]$$

which means that the defense attorney will not strike when $\pi_1 = \overline{g}$ or $\overline{G}$ and will strike when $\pi_1 = \underline{g}$. If $\pi_1 = \underline{G}$, the defense attorney will strike juror 1 only if

$$\mathbb{P}[s_1 > s_1^* | \omega, \pi_1 = \underline{G}] > \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{g}] + \frac{1}{2}\mathbb{P}[s_3 > s_3^* | \omega, g_3 = G]$$

and otherwise will not strike juror 1.

The prosecutor will strike juror 1 if

$$\mathbb{P}[s_1 > s_1^* | \omega, \pi_1] < \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \overline{g}] + \frac{1}{2}\mathbb{P}[s_3 > s_3^* | \omega, g_3 = G]$$

which means that the prosecutor will not strike when $\pi_1 = \underline{g}$ or $\underline{G}$ and will strike when $\pi_1 = \overline{g}$. If $\pi_1 = \overline{G}$, the prosecutor will strike juror 1 only if

$$\mathbb{P}[s_1 > s_1^* | \omega, \pi_1 = \overline{G}] < \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \overline{g}] + \frac{1}{2}\mathbb{P}[s_3 > s_3^* | \omega, g_3 = G]$$

and otherwise will not strike juror 1.

So, we have the following distributions for different values of π_1: when $\pi_1 = \underline{g}$ (defense strikes),

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
$\frac{1}{2}$	0	$\frac{1}{4}$	$\frac{1}{4}$

when $\pi_1 = \overline{g}$ (prosecutor strikes),

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
0	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

when $\pi_1 = \underline{G}$ and $\mathbb{P}[s_1 > s_1^* | \omega, \pi_1 = \underline{G}] > \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{g}] + \frac{1}{2}\mathbb{P}[s_3 > s_3^* | \omega, g_3 = G]$ (defense strikes),

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
$\frac{1}{2}$	0	$\frac{1}{4}$	$\frac{1}{4}$

when $\pi_1 = \underline{G}$ and $\mathbb{P}[s_1 > s_1^* | \omega, \pi_1 = \underline{G}] \leq \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{g}] + \frac{1}{2}\mathbb{P}[s_3 > s_3^* | \omega, g_3 = G]$ (no strikes),

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
0	0	1	0

when $\pi_1 = \overline{G}$ and $\mathbb{P}[s_1 > s_1^* | \omega, \pi_1 = \overline{G}] \geq \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \overline{g}] + \frac{1}{2}\mathbb{P}[s_3 > s_3^* | \omega, g_3 = G]$ (no strikes),

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
0	0	0	1

and when $\pi_1 = \overline{G}$ and $\mathbb{P}[s_1 > s_1^* | \omega, \pi_1 = \overline{G}] < \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \overline{g}] + \frac{1}{2}\mathbb{P}[s_3 > s_3^* | \omega, g_3 = G]$ (prosecutor strikes),

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
0	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

Case 3: $\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{g}] \geq \mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \overline{g}] \geq \mathbb{P}[s_3 > s_3^* | \omega, g_3 = G]$

In this case, when choosing between jurors 2 and 3, the defense attorney will always choose to strike juror 2 and the prosecutor will always choose to keep juror 2. So, the defense attorney will strike juror 1 if

$$1 - \mathbb{P}[s_1 > s_1^* | \omega, \pi_1] < 1 - \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{g}] - \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_3 = \overline{g}]$$

which means that the defense attorney will not strike when $\pi_1 = \overline{g}$ or $\overline{G}$ and will strike when $\pi_1 = \underline{g}$. If $\pi_1 = \underline{G}$, then the defense attorney only strikes if $\mathbb{P}[s_1 > s_1^* | \omega, \pi_1 = \underline{G}] > \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{g}] + \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_3 = \overline{g}]$ and otherwise does not strike.

The prosecutor will strike juror 1 if

$$\mathbb{P}[s_1 > s_1^* | \omega, \pi_1] < \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{G}] + \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \overline{G}]$$

which means that the prosecutor will not strike when $\pi_1 = \underline{g}, \overline{g}$, or $\underline{G}$ and will strike only when $\pi_1 = \overline{G}$.

So, we have the following distributions for different values of π_1: when $\pi_1 = \underline{g}$ (defense strikes),

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
$\frac{1}{2}$	$\frac{1}{2}$	0	0

when $\pi_1 = \overline{g}$ (no strikes),

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
0	1	0	0

when $\pi_1 = \underline{G}$ and $\mathbb{P}[s_1 > s_1^* | \omega, \pi_1 = \underline{G}] > \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{g}] + \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_3 = \overline{g}]$ (defense strikes),

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
$\frac{1}{2}$	$\frac{1}{2}$	0	0

when $\pi_1 = \underline{G}$ and $\mathbb{P}[s_1 > s_1^* | \omega, \pi_1 = \underline{G}] \leq \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_2 = \underline{g}] + \frac{1}{2}\mathbb{P}[s_2 > s_2^* | \omega, \pi_3 = \overline{g}]$ (no strikes),

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
0	0	1	0

and when $\pi_1 = \overline{G}$ (prosecutor strikes),

$\underline{g}$	$\overline{g}$	$\underline{G}$	$\overline{G}$
0	0	$\frac{1}{2}$	$\frac{1}{2}$

Appendix C

SPECIAL CASE: DERIVATION OF JUROR DISTRIBUTIONS

The special case in which group B has a single value b for π_i simplifies the attorney behavior calculations because knowing that a juror is from group B fully determines the juror's cutoff s_i^* for conviction. Then, the attorney behavior determines the juror distributions (the probability that the selected juror has each of the possible payoff values: $\underline{a}, \overline{a}, b$). Considering each possible sequence of groups for the jury pool yields the following results.

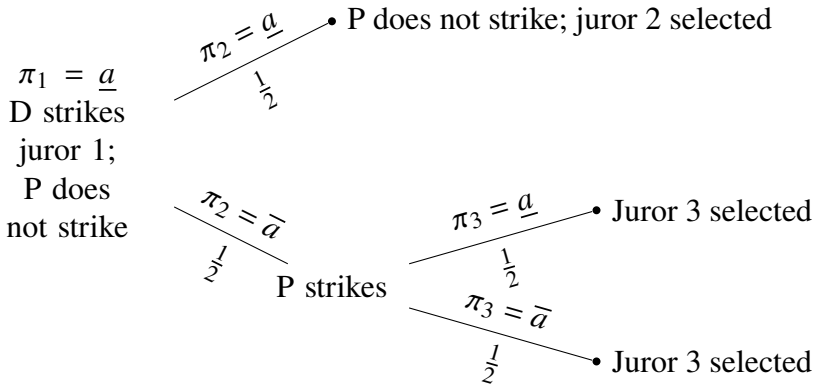
AAA

When all three jurors in the pool come from group A, the attorney for whom π_1 is worse will strike juror 1.

So, when $\pi_1 = \underline{a}$, the selected juror will have $\pi_i = \underline{a}$ three-quarters of the time and $\pi_i = \overline{a}$ one-quarter of the time. When $\pi_1 = \overline{a}$, the selected juror will have $\pi_i = \overline{a}$ three-quarters of the time and $\pi_i = \underline{a}$ one-quarter of the time. Since $\pi_1 = \underline{a}$ with probability $\frac{1}{2}$, the juror distribution for this sequence of jury pool groups is

$\underline{a}$	$\overline{a}$	b
$\frac{1}{2}$	$\frac{1}{2}$	0

Figure C.1: Attorney Strike Choices for Group Sequence AAA



AAB

The juror distributions here will depend on the probability of conviction given that the juror comes from group B relative to the probabilities of conviction given that the juror has $\pi_i = \underline{a}$ and $\overline{a}$. In this special case, the probability of conviction from a group B juror equals the probability of conviction from a juror with $\pi_i = b$.

If $c(\omega, b) > c(\omega, \underline{a}) \geq c(\omega, \overline{a})$, the defense attorney will never strike the first juror because it would give the prosecutor the option to strike the second juror and have a juror with the highest known probability of conviction. Similarly, if the prosecutor strikes the first juror, the defense attorney will never strike the second juror. So, if $\pi_1 = \underline{a}$ no one will strike and if $\pi_1 = \overline{a}$ the prosecutor will strike and the defense will not, giving equal chances of the selected juror having $\pi_2 = \underline{a}$ and $\overline{a}$. These behaviors yield the following juror distribution:

$\underline{a}$	$\overline{a}$	b
$\frac{3}{4}$	$\frac{1}{4}$	0

If $c(\omega, \underline{a}) \geq c(\omega, b) > c(\omega, \overline{a})$, the defense attorney will strike juror 1 if $\pi_1 = \underline{a}$ and the prosecutor will strike juror 1 if $\pi_1 = \overline{a}$. The attorney with the remaining strike will keep juror 2 half the time (if π_2 is the better value for that attorney) and otherwise will strike juror 2.

$\underline{a}$	$\overline{a}$	b
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$

If $c(\omega, \underline{a}) \geq c(\omega, \overline{a}) \geq c(\omega, b)$, the prosecutor never strikes the first juror or the second juror, and the defense attorney only strikes when $\pi_1 = \underline{a}$.

$\underline{a}$	$\overline{a}$	b
$\frac{1}{4}$	$\frac{3}{4}$	0

ABA

If $c(\omega, b) > c(\omega, \underline{a})$, then the defense would always strike juror 2 and the prosecutor would always keep juror 2 when given the chance. Then, when $\pi_1 = \underline{a}$, neither

attorney will strike juror 1. When $\pi_1 = \overline{a}$, the prosecutor will strike juror 1 and the defense will strike juror 2.

$\underline{a}$	$\overline{a}$	b
$\frac{3}{4}$	$\frac{1}{4}$	0

If $c(\omega, \underline{a}) > c(\omega, b) > c(\omega, A)$, then the defense would always strike juror 2 and the prosecutor would always keep juror 2 when given the chance. When $\pi_1 = \underline{a}$, the defense would strike juror 1, and the prosecutor would keep juror 2. When $\pi_1 = \overline{a}$, the prosecutor would strike juror 1 and the defense would strike juror 2.

$\underline{a}$	$\overline{a}$	b
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$

If $c(\omega, A) > c(\omega, b) > c(\omega, \overline{a})$, then the defense will keep juror 2 and the prosecutor will strike juror 2. When $\pi_1 = \underline{a}$, the defense would strike juror 1, and the prosecutor would strike juror 2. When $\pi_1 = \overline{a}$, the prosecutor would strike juror 1, and the defense would keep juror 2.

$\underline{a}$	$\overline{a}$	b
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$

If $c(\omega, \overline{a}) > c(\omega, b)$, then the defense will keep juror 2 and the prosecutor will strike juror 2. When $\pi_1 = \underline{a}$, the defense would strike juror 1, and the prosecutor would strike juror 2. When $\pi_1 = \overline{a}$, neither attorney will strike juror 1.

$\underline{a}$	$\overline{a}$	b
$\frac{1}{4}$	$\frac{3}{4}$	0

BAA

In this case, the effect of striking the first juror will be the same as for striking the first juror in the AAA case: the payoff in $\{\underline{a}, \overline{a}\}$ better for the other attorney will occur $\frac{3}{4}$ of the time.

If $c(\omega, b) > \frac{3}{4}c(\omega, \underline{a}) + \frac{1}{4}c(\omega, \overline{a})$, then the defense always will strike juror 1, and the prosecutor will strike juror 2 only if $\pi_2 = \overline{a}$.

$\underline{a}$	$\overline{a}$	b
$\frac{3}{4}$	$\frac{1}{4}$	0

If $\frac{3}{4}c(\omega, \underline{a}) + \frac{1}{4}c(\omega, \overline{a}) \geq c(\omega, b) \geq \frac{1}{4}c(\omega, \underline{a}) + \frac{3}{4}c(\omega, \overline{a})$, then neither attorney will strike juror 1.

$\underline{a}$	$\overline{a}$	b
0	0	1

If $\frac{1}{4}c(\omega, \underline{a}) + \frac{3}{4}c(\omega, \overline{a}) > c(\omega, b)$, then the prosecutor will strike juror 1, and the defense will strike juror 2 only if $\pi_2 = \underline{a}$.

$\underline{a}$	$\overline{a}$	b
$\frac{1}{4}$	$\frac{3}{4}$	0

ABB

Here, the outcome of striking juror 1 will always be a juror with $\pi_i = b$. That outcome will be better than the revealed π_1 for one of the attorneys, and so juror 1 will always be struck.

$\underline{a}$	$\overline{a}$	b
0	0	1

BAB

If $c(\omega, b) > c(\omega, \underline{a})$, then the defense would keep juror 2, and the prosecutor would strike juror 2. The prosecutor will not strike juror 1, and so the defense is indifferent between striking and not and the selected juror will always be from group B.

If $c(\omega, \underline{a}) > c(\omega, b) > c(\omega, \overline{a})$, then the defense will strike juror 2 if $\pi_2 = \underline{a}$ and the prosecutor will strike juror 2 if $\pi_2 = \overline{a}$. As a result, neither attorney will strike juror 1.

If $c(\omega, \overline{a}) > c(\omega, b)$, then the defense would strike juror 2, and the prosecutor would keep juror 2. The defense will not strike juror 1, and so the prosecutor is indifferent between striking and not. So, in all cases, the juror distribution is

$\underline{a}$	$\overline{a}$	b
0	0	1

BBA

Here, if the first juror is struck, the attorney with the remaining strike will gain no new information, as π_2 was already known. The attorney for whom $\pi_i = b$ is worse than the expected probability of conviction for juror 3 from group A knows that the other attorney would keep juror 2 if juror 1 is struck, and so would be indifferent between striking juror 1 and not. The selected juror will always be from group B.

$\underline{a}$	$\overline{a}$	b
0	0	1

BBB

Here, the selected juror will always be from B and so will have $\pi_i = b$.

$\underline{a}$	$\overline{a}$	b
0	0	1

Appendix D

COMPARISON TO OTHER JURY SELECTION PROCEDURES

Section 1.4 compares a random draw to a peremptory strike procedure where the attorneys can learn juror type through questioning and can observe the group memberships of upcoming jurors. The random draw reflects a system that has eliminated peremptory strikes. The following sections consider several other institutional designs.

No Learning, Known Group Sequence (Pure Stereotyping with Known Juror Order)

The full revelation of type in the base model assumes that attorneys can gather significant information from questioning potential jurors, beyond their group affiliation. However, some courts' procedures give little opportunity for attorneys to examine the jury, and some attorneys may be less skilled at extracting information through questioning (see Lehmann & Smith, 2013). This variant of the model reflects those circumstances.

Here, the attorneys know the group sequence for the three potential jurors, but they can never learn more information about any of them. As such, attorneys base all strike decisions on the expected probabilities of conviction for the groups: $c(\omega, A)$ and $c(\omega, B) = c(\omega, b)$. One attorney will prefer group A and the other group B. The group sequence will contain more jurors from A or more jurors from B. Whichever attorney prefers the group that appears more often in the sequence will choose strikes to force the selected juror to come from that group — if the worse group appears first, the attorney will strike the first juror and both remaining jurors will be from the better group; if the better group appears first, the attorney will save her strike, and if the opponent strikes the first juror, the attorney can strike or not strike the second juror to guarantee selecting the better group. Since this process reveals no type information, the attorneys cannot condition their strikes on type. The juror chosen will be a random choice from the juror's group.

So, for group sequences AAA, AAB, ABA, and BAA, a random member of group A will become the juror. For group sequences ABB, BAB, BBA, and BBB, a random member of group B will become the juror. Then, the probability of selecting an $\underline{a}$

juror equals the probability of selecting a $\overline{a}$ juror:

$$\frac{1}{2}\alpha^3 + \frac{3}{2}\alpha^2(1-\alpha)$$

The probability of selecting a b juror is

$$3\alpha(1-\alpha)^2 + (1-\alpha)^3$$

These juror distributions imply Proposition 8.

Proposition 8. *For a zero-variance minority, when attorneys can observe only the group memberships of the sequence of jurors and cannot learn about their types,*

1. *the majority is overrepresented on the jury,*
2. *the distribution of types of majority jurors is the same as the distribution of types of majority group members,*
3. *the selected jury's expected probability of conviction (and of false convictions) is above (below) that of a random draw from the population if the expected probability of conviction of a minority juror is below (above) that of an average majority juror; the expected probability of false acquittals is below (above) that of a random draw from the population if the expected probability of conviction of a minority juror is below (above) that of an average majority juror.*

Learning Only about Groups, No Knowledge of Sequence (Pure Stereotyping with Unknown Juror Order)

This variant involves attorneys who only can stereotype by groups and who also know nothing about the juror sequence — unlike the prior variant, they do not know the group memberships of upcoming potential jurors.

If $c(\omega, B) < c(\omega, A)$, then the defense will strike juror 2 if $g_2 = A$ and the prosecutor if $g_2 = B$. So, the defense compares $c(\omega, g_1)$ to $\alpha c(\omega, A) + (1-\alpha)c_{pop}(\omega)$, where $c_{pop}(\omega) \equiv \alpha c(\omega, A) + (1-\alpha)c(\omega, B)$ is the probability of conviction from a random draw from the population. The prosecutor compares $c(\omega, g_1)$ to $\alpha c_{pop}(\omega) + (1-\alpha)c(\omega, B)$. So, the defense will strike a juror 1 from group A and the prosecutor will strike a juror 1 from group B.

Note that the strike decisions in the prior variant (when the group sequence is known) are equivalent to both attorneys always striking the group that is worse for them. So, the results of this variant are exactly the same as for the prior one.

Proposition 9. *For a zero-variance minority, when attorneys learn only juror group memberships through questioning and know nothing about the sequence, attorney behaviors are exactly the same as when the attorneys know the group sequence. The results of this variant are identical to the results in Proposition 8.*

Learning about Types, No Knowledge of Sequence (Effective Questioning with Unknown Juror Order)

This variant removes the attorneys' knowledge of the group sequence. The resulting process is more like a standard secretary problem, or a strike-and-replace system in which the attorneys cannot see the jury pool order.

When deciding whether to strike juror 2, each attorney will compare the revealed juror's probability of conviction $c(\omega, \pi_2)$ with the probability of conviction from a random draw from the population:

$$\begin{aligned} c_{pop}(\omega) &\equiv \frac{\alpha}{2}c(\omega, \underline{a}) + \frac{\alpha}{2}c(\omega, \overline{a}) + (1-\alpha)c(\omega, b) \\ &= \alpha c(\omega, A) + (1-\alpha)c(\omega, b) \end{aligned}$$

Proposition 10. *For a zero-variance minority, when attorneys learn juror types through questioning and know nothing about the sequence,*

1. *when* $c(\omega, \underline{a}) - c(\omega, \overline{a}) < \frac{2(1-\alpha)^2}{2\alpha-\alpha^2}[c(\omega, \overline{a}) - c(\omega, b)]$,
   a) *the majority is overrepresented on the jury,*
   b) *the distribution of types of majority jurors is the same as the distribution of types of majority group members, and*
   c) *the selected jury's expected probability of conviction is above that of a random draw from the population;*
2. *when* $\frac{2(1-\alpha)^2}{2\alpha-\alpha^2}[c(\omega, \overline{a}) - c(\omega, b)] < c(\omega, \underline{a}) - c(\omega, \overline{a}) < \frac{2(1-\alpha)}{\alpha}[c(\omega, \overline{a}) - c(\omega, b)]$,
   a) *the majority is overrepresented on the jury (even more than in the prior case),*
   b) *the distribution of types of majority jurors skews towards* $\overline{a}$*, and*
   c) *the selected jury's expected probability of conviction is above that of a random draw from the population;*
3. *when* $\frac{2(1-\alpha)}{\alpha}[c(\omega, \overline{a}) - c(\omega, b)] < c(\omega, \underline{a}) - c(\omega, \overline{a})$, $c(\omega, b) < c(\omega, A)$, *and*

a) $\frac{2+2\alpha}{\alpha}[c(\omega, b) - c(\omega, \overline{a})] < c(\omega, \underline{a}) - c(\omega, \overline{a}) < \frac{4-2\alpha-2\alpha^2}{\alpha^2}[c(\omega, \overline{a}) - c(\omega, b)]$,

   i. *the majority is overrepresented on the jury,*

   ii. *the distribution of types of majority jurors skews towards* $\overline{a}$*, and*

   iii. *the selected jury's expected probability of conviction is above that of a random draw from the population for lower values of* $c(\omega, b)$ *— i.e., when* $c(\omega, \underline{a}) - c(\omega, \overline{a}) < \frac{6\alpha^2-6\alpha^3}{4\alpha-6\alpha^2+3\alpha^3}[c(\omega, \overline{a}) - c(\omega, b)]$ *— and is below that of a random draw for higher values of* $c(\omega, b)$*;*

b) $\frac{2+2\alpha}{\alpha}[c(\omega, b) - c(\omega, \overline{a})] < c(\omega, \underline{a}) - c(\omega, \overline{a})$ *and* $\frac{4-2\alpha-2\alpha^2}{\alpha^2}[c(\omega, \overline{a}) - c(\omega, b)] < c(\omega, \underline{a}) - c(\omega, \overline{a})$,

   i. *the majority is underrepresented on the jury,*

   ii. *the distribution of types of majority jurors skews towards* $\overline{a}$*, and*

   iii. *the selected jury's expected probability of conviction is below that of a random draw from the population;*

c) $\frac{4-2\alpha-2\alpha^2}{\alpha^2}[c(\omega, \overline{a}) - c(\omega, b)] < c(\omega, \underline{a}) - c(\omega, \overline{a}) < \frac{2+2\alpha}{\alpha}[c(\omega, b) - c(\omega, \overline{a})]$,

   i. *the majority is underrepresented on the jury,*

   ii. *the distribution of types of majority jurors is the same as the distribution of types of majority group members, and*

   iii. *the selected jury's expected probability of conviction is below that of a random draw from the population; and*

4. *the results for* $c(\omega, b) > c(\omega, A)$ *mirror the results above for* $c(\omega, b) < c(\omega, A)$.

Note that the cases in this proposition are ordered so that $c(\omega, b)$ increases for some fixed $c(\omega, \underline{a}$ and $c(\omega, \overline{a})$. Appendix E contains the derivation of these results.

Tables D.1 and D.2 summarize these institutional results.

Table D.1: Zero-Variance Minority: Institutional Comparison of Juror Distribution, Majority Representation, and Probability of Conviction.

	$c(\omega, b)$ is in the interval:	$(0, c(\omega, \overline{a})]$	$[c(\omega, \overline{a}),$ $\frac{1}{4}c(\omega, \underline{a}) + \frac{3}{4}c(\omega, \overline{a})]$	$[\frac{1}{4}c(\omega, \underline{a}) + \frac{3}{4}c(\omega, \overline{a}),$ $c(\omega, A)]$	$[c(\omega, A),$ $\frac{3}{4}c(\omega, \underline{a}) + \frac{1}{4}c(\omega, \overline{a})]$	$[\frac{3}{4}c(\omega, \underline{a}) + \frac{1}{4}c(\omega, \overline{a}),$ $c(\omega, \underline{a})]$	$[c(\omega, \underline{a}), 1)$
Only Group Information	$\mathbb{P}[\pi = \underline{a}]$	$\frac{1}{2}\alpha^3 + \frac{3}{2}\alpha^2(1-\alpha)$					
	$\mathbb{P}[\pi = \overline{a}]$	$\frac{1}{2}\alpha^3 + \frac{3}{2}\alpha^2(1-\alpha)$					
	$\mathbb{P}[\pi = b]$	$3\alpha(1-\alpha)^2 + (1-\alpha)^3$					
	Majority Group A	Overrepresented					
	Conviction Rate vs. Random	Above			Below		
Known Type Sequence	$\mathbb{P}[\pi = \underline{a}]$	$\frac{1}{2}\alpha^3 + \frac{3}{4}\alpha^2(1-\alpha)$					$\frac{1}{2}\alpha^3 + \frac{9}{4}\alpha^2(1-\alpha)$
	$\mathbb{P}[\pi = \overline{a}]$	$\frac{1}{2}\alpha^3 + \frac{9}{4}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{3}{4}\alpha^2(1-\alpha)$				
	$\mathbb{P}[\pi = b]$	$3\alpha(1-\alpha)^2 + (1-\alpha)^3$	$\frac{3}{2}\alpha^2(1-\alpha) + 3\alpha(1-\alpha)^2 + (1-\alpha)^3$				$3\alpha(1-\alpha)^2 + (1-\alpha)^3$
	Majority Group A	Overrepresented	Underrepresented				Overrepresented
	Conviction Rate vs. Random	Above for lower $c(\omega, b)$; Below for higher	Below		Above		Above for lower $c(\omega, b)$; Below for higher
Base Model	$\mathbb{P}[\pi = \underline{a}]$	$\frac{1}{2}\alpha^3 + \frac{3}{4}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{3}{4}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{1}{2}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{1}{2}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{5}{4}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{9}{4}\alpha^2(1-\alpha)$
	$\mathbb{P}[\pi = \overline{a}]$	$\frac{1}{2}\alpha^3 + \frac{9}{4}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{5}{4}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{1}{2}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{1}{2}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{3}{4}\alpha^2(1-\alpha)$	$\frac{1}{2}\alpha^3 + \frac{3}{4}\alpha^2(1-\alpha)$
	$\mathbb{P}[\pi = b]$	$3\alpha(1-\alpha)^2 + (1-\alpha)^3$	$\alpha^2(1-\alpha)+$ $3\alpha(1-\alpha)^2 + (1-\alpha)^3$	$2\alpha^2(1-\alpha)+$ $3\alpha(1-\alpha)^2 + (1-\alpha)^3$	$2\alpha^2(1-\alpha)+$ $3\alpha(1-\alpha)^2 + (1-\alpha)^3$	$\alpha^2(1-\alpha)+$ $3\alpha(1-\alpha)^2 + (1-\alpha)^3$	$3\alpha(1-\alpha)^2 + (1-\alpha)^3$
	Majority Group A	Overrepresented	Underrepresented				Overrepresented
	Conviction Rate vs. Random	Above for lower $c(\omega, b)$; Below for higher	Below		Above		Above for lower $c(\omega, b)$; Below for higher

Table D.2: Zero-Variance Minority: Juror Distribution, Majority Representation, and Probability of Conviction for Learning with No Knowledge of Sequence; Columns Ordered by Increasing Values of $c(\omega, b)$ Given Fixed $c(\omega, \underline{a})$ and $c(\omega, \underline{a})$; See Appendix E for Boundary Values.

Learn Type, No Seq. Known										
	$\mathbb{P}[\pi = \underline{a}]$	$\frac{3}{2}\alpha^2 - \alpha^3$	$\alpha^2 - \frac{3}{4}\alpha^3$	$\frac{3}{4}\alpha^2 - \frac{3}{8}\alpha^3$	$\frac{3}{4}\alpha^2 - \frac{1}{4}\alpha^3$	$\frac{1}{2}\alpha^2$	$\frac{1}{2}\alpha + \frac{1}{4}\alpha^2 - \frac{1}{4}\alpha^3$	$\alpha - \frac{3}{8}\alpha^3$	$\frac{1}{2}\alpha + \alpha^2 - \frac{3}{4}\alpha^3$	$\frac{3}{2}\alpha^2 - \alpha^3$
	$\mathbb{P}[\pi = \overline{a}]$	$\frac{3}{2}\alpha^2 - \alpha^3$	$\frac{1}{2}\alpha + \alpha^2 - \frac{3}{4}\alpha^3$	$\alpha - \frac{3}{8}\alpha^3$	$\frac{1}{2}\alpha + \frac{1}{4}\alpha^2 - \frac{1}{4}\alpha^3$	$\frac{1}{2}\alpha^2$	$\frac{3}{4}\alpha^2 - \frac{1}{4}\alpha^3$	$\frac{3}{4}\alpha^2 - \frac{3}{8}\alpha^3$	$\alpha^2 - \frac{3}{4}\alpha^3$	$\frac{3}{2}\alpha^2 - \alpha^3$
	$\mathbb{P}[\pi = b]$	$(1-\alpha)^2 + 2\alpha(1-\alpha)^2$	$(1-\alpha)^2 + \frac{3}{2}\alpha(1-\alpha)^2$	$1-\alpha-\frac{3}{4}\alpha^2+\frac{3}{4}\alpha^3$	$1-\frac{1}{2}\alpha-\alpha^2+\frac{1}{2}\alpha^3$	$1-\alpha^2$	$1-\frac{1}{2}\alpha-\alpha^2+\frac{1}{2}\alpha^3$	$1-\alpha-\frac{3}{4}\alpha^2+\frac{3}{4}\alpha^3$	$(1-\alpha)^2 + \frac{3}{2}\alpha(1-\alpha)^2$	$(1-\alpha)^2 + 2\alpha(1-\alpha)^2$
	Majority Group A	Overrepresented			Underrepresented			Overrepresented		
	Conviction Rate vs. Random	Above		Above for lower $c(\omega, b)$; Below for higher	Below		Above	Above for lower $c(\omega, b)$; Below for higher		Below

Appendix E

JUROR DISTRIBUTIONS AND PROBABILITY OF CONVICTION FOR LEARNING ABOUT TYPES, NO KNOWLEDGE OF SEQUENCE (EFFECTIVE QUESTIONING WITH UNKNOWN JUROR ORDER)

When deciding whether to strike juror 2, each attorney will compare the revealed juror's probability of conviction $c(\omega, \pi_2)$ with the probability of conviction from a random draw from the population:

$$\begin{aligned} c_{pop}(\omega) &\equiv \frac{\alpha}{2}c(\omega, \underline{a}) + \frac{\alpha}{2}c(\omega, \overline{a}) + (1-\alpha)c(\omega, b) \\ &= \alpha c(\omega, A) + (1-\alpha)c(\omega, b) \end{aligned}$$

Then, the attorneys' strike behaviors change across four regions, as the value of $c(\omega, b)$ varies.

$c(\omega, b) < \dfrac{c(\omega, \overline{a}) - \alpha c(\omega, A)}{1-\alpha}$

Note that this condition is equivalent to

$$c(\omega, \underline{a}) - c(\omega, \overline{a}) < \frac{2(1-\alpha)}{\alpha}[c(\omega, \overline{a}) - c(\omega, b)]$$

and the coefficient $\dfrac{2(1-\alpha)}{\alpha}$ decreases from 2 to 0 as α increases from $\frac{1}{2}$ to 1. Note also that this condition can hold only if $c(\omega, b) < c(\omega, \overline{a})$.

In this region,

$$\begin{aligned} c(\omega, \underline{a}) &> c_{pop}(\omega) \\ c(\omega, \overline{a}) &> c_{pop}(\omega) \\ c(\omega, b) &< c_{pop}(\omega) \end{aligned}$$

and the defense would strike a juror 2 with $\pi_2 = \underline{a}$ or $\overline{a}$, while the prosecutor would strike a juror 2 with $\pi_2 = b$. Then the defense would compare juror 1 to

$$\frac{\alpha}{2}c(\omega, \underline{a}) + \frac{\alpha}{2}c(\omega, \overline{a}) + (1-\alpha)c_{pop}(\omega)$$

and the prosecutor would compare juror 1 to

$$\alpha c_{pop}(\omega) + (1-\alpha)c(\omega, b)$$

So, the prosecutor always strikes juror 1 if $\pi_1 = b$.

If

$$c(\omega, b) < \frac{c(\omega, \overline{a}) - [\alpha + \alpha(1-\alpha)]c(\omega, A)}{(1-\alpha)^2}$$

or equivalently

$$c(\omega, \underline{a}) - c(\omega, \overline{a}) < \frac{2(1-\alpha)^2}{2\alpha - \alpha^2}[c(\omega, \overline{a}) - c(\omega, b)]$$

then a juror with $\pi_i = \overline{a}$ has a higher probability of conviction than the prosecutor's choice between jurors 2 and 3. So, the defense would strike a juror 1 with $\pi_1 = \underline{a}$ or $\overline{a}$, and the first juror always gets struck. Note that the coefficient $\frac{2(1-\alpha)^2}{2\alpha - \alpha^2}$ is always smaller than $\frac{2(1-\alpha)}{\alpha}$ so values satisfying this condition will always fall within the region. Also, $\frac{2(1-\alpha)^2}{2\alpha - \alpha^2}$ decreases from $\frac{2}{3}$ to 0 as α increases from $\frac{1}{2}$ to 1.

The attorney strike decisions yield an overall probability of conviction of

$$\begin{aligned}
&\alpha\left[\frac{\alpha}{2}c(\omega, \underline{a}) + \frac{\alpha}{2}c(\omega, \overline{a}) + (1-\alpha)c_{pop}(\omega)\right] + (1-\alpha)\left[\alpha c_{pop}(\omega) + (1-\alpha)c(\omega, b)\right] \\
&\quad = \left(\frac{3}{2}\alpha^2 - \alpha^3\right)c(\omega, \underline{a}) + \left(\frac{3}{2}\alpha^2 - \alpha^3\right)c(\omega, \overline{a}) + \left[(1-\alpha)^2 + 2\alpha(1-\alpha)^2\right]c(\omega, b)
\end{aligned}$$

So, the majority group A is overrepresented, and a majority juror has the same type distribution as a random draw from A. As such, the probability of conviction is above the probability of conviction from a random draw.

If

$$\frac{2(1-\alpha)^2}{2\alpha - \alpha^2}[c(\omega, \overline{a}) - c(\omega, b)] < c(\omega, \underline{a}) - c(\omega, \overline{a}) < \frac{2(1-\alpha)}{\alpha}[c(\omega, \overline{a}) - c(\omega, b)]$$

then the defense would only strike juror 1 if $\pi_1 = \underline{a}$.

This yields an overall probability of conviction of

$$\begin{aligned}
&\tfrac{\alpha}{2}\left[\tfrac{\alpha}{2}c(\omega, \underline{a}) + \tfrac{\alpha}{2}c(\omega, \overline{a}) + (1-\alpha)c_{pop}(\omega)\right] + \tfrac{\alpha}{2}c(\omega, \overline{a}) + (1-\alpha)\left[\alpha c_{pop}(\omega) + (1-\alpha)c(\omega, b)\right] \\
&\quad = \left(\alpha^2 - \tfrac{3}{4}\alpha^3\right)c(\omega, \underline{a}) + \left(\tfrac{1}{2}\alpha + \alpha^2 - \tfrac{3}{4}\alpha^3\right)c(\omega, \overline{a}) + \left[(1-\alpha)^2 + \tfrac{3}{2}\alpha(1-\alpha)^2\right]c(\omega, b)
\end{aligned}$$

In this subregion, A is even more overrepresented, but the distribution of A jurors skews towards $\overline{a}$, the type closer to b. Since the expected probability of conviction from either type of A juror is above the expected probability of conviction for the population, the overrepresentation of A implies that the jury is more likely to convict than a randomly drawn juror.

$$\frac{c(\omega,\overline{a}) - \alpha c(\omega, A)}{1-\alpha} < c(\omega, b) < c(\omega, A)$$

In this region,

$$c(\omega,\underline{a}) > c_{pop}(\omega)$$
$$c(\omega,\overline{a}) < c_{pop}(\omega)$$
$$c(\omega, b) < c_{pop}(\omega)$$

So, the defense would strike juror 2 if $\pi_2 = \underline{a}$ and the prosecutor would strike if $\pi_2 = \overline{a}$ or b.

The defense compares juror 1 to

$$\frac{\alpha}{2}c(\omega,\underline{a}) + \left(1 - \frac{\alpha}{2}\right) c_{pop}(\omega)$$

and will strike if $\pi_1 = \underline{a}$.

The prosecutor compares juror 1 to

$$\frac{\alpha}{2}c_{pop}(\omega) + \frac{\alpha}{2}c(\omega,\overline{a}) + (1-\alpha)c(\omega, b)$$

The prosecutor will strike $\pi_1 = b$ if $c(\omega, b)$ is less than that value or equivalently when

$$c(\omega,\underline{a}) - c(\omega,\overline{a}) > \frac{2+2\alpha}{\alpha}[c(\omega, b) - c(\omega,\overline{a})]$$

The coefficient $\frac{2+2\alpha}{\alpha}$ decreases from 6 to 4 as α increases from $\frac{1}{2}$ to 1. Note that this inequality always holds when $c(\omega, b) < c(\omega,\overline{a})$ and also holds for some region of values of $c(\omega, b)$ above $c(\omega,\overline{a})$.

The prosecutor will strike $\pi_1 = \overline{a}$ if $c(\omega,\overline{a})$ is less than the expected probability of conviction from the defense choosing whether to strike juror 2 or equivalently if

$$c(\omega,\underline{a}) - c(\omega,\overline{a}) > \frac{4-2\alpha-2\alpha^2}{\alpha^2}[c(\omega,\overline{a}) - c(\omega, b)]$$

The coefficient $\frac{4-2\alpha-2\alpha^2}{\alpha^2}$ decreases from 10 to 0 as α increases from $\frac{1}{2}$ to 1. Note that this inequality always holds when $c(\omega, b) > c(\omega,\overline{a})$ and also holds for some region of values of $c(\omega, b)$ below $c(\omega,\overline{a})$.

So, if only the first condition on $c(\omega,\underline{a}) - c(\omega,\overline{a})$ holds, the defense will strike $\pi_1 = \underline{a}$, neither will strike $\pi_1 = \overline{a}$, and the prosecutor will strike $\pi_1 = b$. The expected probability of conviction will be

$$\frac{\alpha}{2}\left[\frac{\alpha}{2}c(\omega,\underline{a}) + \left(1-\frac{\alpha}{2}\right)c_{pop}(\omega)\right] + \frac{\alpha}{2}c(\omega,\overline{a}) + (1-\alpha)\left[\frac{\alpha}{2}c_{pop}(\omega) + \frac{\alpha}{2}c(\omega,\overline{a}) + (1-\alpha)c(\omega, b)\right]$$
$$= \left(\frac{3}{4}\alpha^2 - \frac{3}{8}\alpha^3\right)c(\omega,\underline{a}) + \left(\alpha - \frac{3}{8}\alpha^3\right)c(\omega,\overline{a}) + \left[1 - \alpha - \frac{3}{4}\alpha^2 + \frac{3}{4}\alpha^3\right]c(\omega, b)$$

Here, the majority is overrepresented, and the distribution of the majority juror skews towards $\overline{a}$.

The overall probability of conviction is lower than a random draw when

$$c(\omega, \underline{a}) - c(\omega, \overline{a}) > \frac{6\alpha^2 - 6\alpha^3}{4\alpha - 6\alpha^2 + 3\alpha^3}[c(\omega, \overline{a}) - c(\omega, b)]$$

Note that, given the region considered in this subsection, this condition always holds with a large enough minority: $\alpha < \frac{2}{3}$. For other values of α, the probability of conviction will be higher than a random draw for lower values of $c(\omega, b)$ and lower than a random draw for higher values of $c(\omega, b)$.

When both conditions on $c(\omega, \underline{a}) - c(\omega, \overline{a})$ hold, the defense will strike $\pi_1 = \underline{a}$, and the prosecutor will strike when $\pi_1 = \overline{a}$ or b. The expected probability of conviction will be

$$\begin{aligned}
&\frac{\alpha}{2}\left[\frac{\alpha}{2}c(\omega, \underline{a}) + \left(1 - \frac{\alpha}{2}\right)c_{pop}(\omega)\right] + (1 - \frac{\alpha}{2})\left[\frac{\alpha}{2}c_{pop}(\omega) + \frac{\alpha}{2}c(\omega, \overline{a}) + (1-\alpha)c(\omega, b)\right] \\
&= \left(\frac{3}{4}\alpha^2 - \frac{1}{4}\alpha^3\right)c(\omega, \underline{a}) + \left(\frac{1}{2}\alpha + \frac{1}{4}\alpha^2 - \frac{1}{4}\alpha^3\right)c(\omega, \overline{a}) + \left(1 - \frac{1}{2}\alpha - \alpha^2 + \frac{1}{2}\alpha^3\right)c(\omega, b)
\end{aligned}$$

Here, the majority is underrepresented, and the distribution of the majority juror skews towards $\overline{a}$.

The overall probability of conviction is lower than a random draw when

$$c(\omega, \underline{a}) - c(\omega, \overline{a}) > \frac{2\alpha - 4\alpha^2 + 2\alpha^3}{2\alpha - 3\alpha^2 + \alpha^3}[c(\omega, b) - c(\omega, \overline{a})]$$

The coefficient here is always smaller than the coefficient for the condition that the prosecutor will strike b, and so this condition is always satisfied.

When only the second condition on $c(\omega, \underline{a}) - c(\omega, \overline{a})$ holds, the defense will strike $\pi_1 = \underline{a}$, the prosecutor will strike $\pi_1 = \overline{a}$, and neither will strike $\pi_1 = b$. The expected probability of conviction will be

$$\begin{aligned}
&\frac{\alpha}{2}\left[\frac{\alpha}{2}c(\omega, \underline{a}) + \left(1 - \frac{\alpha}{2}\right)c_{pop}(\omega)\right] + \frac{\alpha}{2}\left[\frac{\alpha}{2}c_{pop}(\omega) + \frac{\alpha}{2}c(\omega, \overline{a}) + (1-\alpha)c(\omega, b)\right] + (1-\alpha)c(\omega, b) \\
&= \left(\frac{1}{2}\alpha^2\right)c(\omega, \underline{a}) + \left(\frac{1}{2}\alpha^2\right)c(\omega, \overline{a}) + \left(1 - \alpha^2\right)c(\omega, b)
\end{aligned}$$

Here, the majority is underrepresented, and the distribution of the majority juror matches the distribution of the majority group. The probability of conviction is lower than a random draw.

$$c(\omega, A) < c(\omega, b) < \frac{c(\omega, \underline{a}) - \alpha c(\omega, A)}{1 - \alpha}$$

In this region,

$$c(\omega, \underline{a}) > c_{pop}(\omega)$$
$$c(\omega, \overline{a}) < c_{pop}(\omega)$$
$$c(\omega, b) > c_{pop}(\omega)$$

The attorney behaviors and juror distributions will be a mirror image of those in the $\frac{c(\omega, \overline{a}) - \alpha c(\omega, A)}{1 - \alpha} < c(\omega, b) < c(\omega, A)$ range.

The prosecutor always strikes $\pi_i = \overline{a}$. The defense strikes $\pi_2 = \underline{a}$ or b. The defense strikes $\pi_2 = \underline{a}$ when

$$c(\omega, \underline{a}) - c(\omega, \overline{a}) > \frac{4 - 2\alpha - 2\alpha^2}{4 - 2\alpha - \alpha^2}[c(\omega, b) - c(\omega, \overline{a})]$$

and strikes $\pi_2 = b$ when

$$c(\omega, \underline{a}) - c(\omega, \overline{a}) < \frac{2\alpha + 2\alpha^2}{2\alpha + \alpha^2}[c(\omega, b) - c(\omega, \overline{a})]$$

When the defense strikes only $\underline{a}$, the expected probability of conviction will be

$$\left(\frac{1}{2}\alpha^2\right) c(\omega, \underline{a}) + \left(\frac{1}{2}\alpha^2\right) c(\omega, \overline{a}) + \left(1 - \alpha^2\right) c(\omega, b)$$

and so the majority will be underrepresented, the distribution of the majority juror matches the distribution of the majority group, and the probability of conviction is higher than a random draw.

When the defense strikes both $\underline{a}$ and b, the expected probability of conviction will be

$$\left(\frac{1}{2}\alpha + \frac{1}{4}\alpha^2 - \frac{1}{4}\alpha^3\right) c(\omega, \underline{a}) + \left(\frac{3}{4}\alpha^2 - \frac{1}{4}\alpha^3\right) c(\omega, \overline{a}) + \left(1 - \frac{1}{2}\alpha - \alpha^2 + \frac{1}{2}\alpha^3\right) c(\omega, b)$$

and so the majority is underrepresented, the distribution of the majority juror skews towards $\underline{a}$, and the overall probability of conviction is higher than a random draw.

When the defense strikes only b, the expected probability of conviction will be

$$\left(\alpha - \frac{3}{8}\alpha^3\right) c(\omega, \underline{a}) + \left(\frac{3}{4}\alpha^2 - \frac{3}{8}\alpha^3\right) c(\omega, \overline{a}) + \left[1 - \alpha - \frac{3}{4}\alpha^2 + \frac{3}{4}\alpha^3\right] c(\omega, b)$$

and so the majority is overrepresented, and the distribution of the majority juror skews towards $\underline{a}$.

The overall probability of conviction is higher than a random draw when

$$c(\omega, \underline{a}) - c(\omega, \overline{a}) > \frac{6\alpha^2 - 6\alpha^3}{4\alpha - 3\alpha^3}[c(\omega, \overline{a}) - c(\omega, b)]$$

The probability of conviction will be higher than a random draw for lower values of $c(\omega, b)$ and lower than a random draw for higher values of $c(\omega, b)$.

$$\frac{c(\omega, \underline{a}) - \alpha c(\omega, A)}{1 - \alpha} < c(\omega, b)$$

In this region,

$$c(\omega, \underline{a}) < c_{pop}(\omega)$$
$$c(\omega, \overline{a}) < c_{pop}(\omega)$$
$$c(\omega, b) > c_{pop}(\omega)$$

The attorney behaviors and juror distributions will be a mirror image of those in the $c(\omega, b) < \frac{c(\omega, \overline{a}) - \alpha c(\omega, A)}{1 - \alpha}$ range.

The defense will always strike when $\pi_i = b$. The prosecution will always strike when $\pi_i = \overline{a}$.

The prosecution will not strike $\pi_2 = \underline{a}$ when

$$c(\omega, \underline{a}) - c(\omega, \overline{a}) > \frac{(1 - \alpha)^2}{1 - \alpha + \frac{1}{2}\alpha^2}[c(\omega, b) - c(\omega, \overline{a})]$$

and the overall probability of conviction will be

$$\left(\frac{1}{2}\alpha + \alpha^2 - \frac{3}{4}\alpha^3\right) c(\omega, \underline{a}) + \left(\alpha^2 - \frac{3}{4}\alpha^3\right) c(\omega, \overline{a}) + \left[(1 - \alpha)^2 + \frac{3}{2}\alpha(1 - \alpha)^2\right] c(\omega, b)$$

meaning that A is overrepresented, the distribution of A jurors skews towards $\underline{a}$, and the jury is less likely to convict than a random draw.

Otherwise, the prosecution will strike $\pi_2 = \underline{a}$, and the overall probability of conviction will be

$$\left(\frac{3}{2}\alpha^2 - \alpha^3\right) c(\omega, \underline{a}) + \left(\frac{3}{2}\alpha^2 - \alpha^3\right) c(\omega, \overline{a}) + \left[(1 - \alpha)^2 + 2\alpha(1 - \alpha)^2\right] c(\omega, b)$$

meaning that A is overrepresented, a majority juror has the same type distribution as a random draw from A, and the jury is less likely to convict than a random draw.

Appendix F

IRT-BASED PROXY AND EMPIRICAL RESULTS

The IRT-based proxy in this appendix uses the same questions as the count-based proxy to find ideal points for pool members using an item response theory model in which the discrimination parameter for a response coded as expressing a bias in favor of the prosecution was constrained to be +5 and the discrimination parameter for a response expressing a bias in favor of the defense was constrained to be −5. The remaining questions had unconstrained parameters.[1] The results are similar to those for the count-based proxy. The IRT-based proxy reflects attorney strike choices, and White pool members tend to be more pro-prosecution:

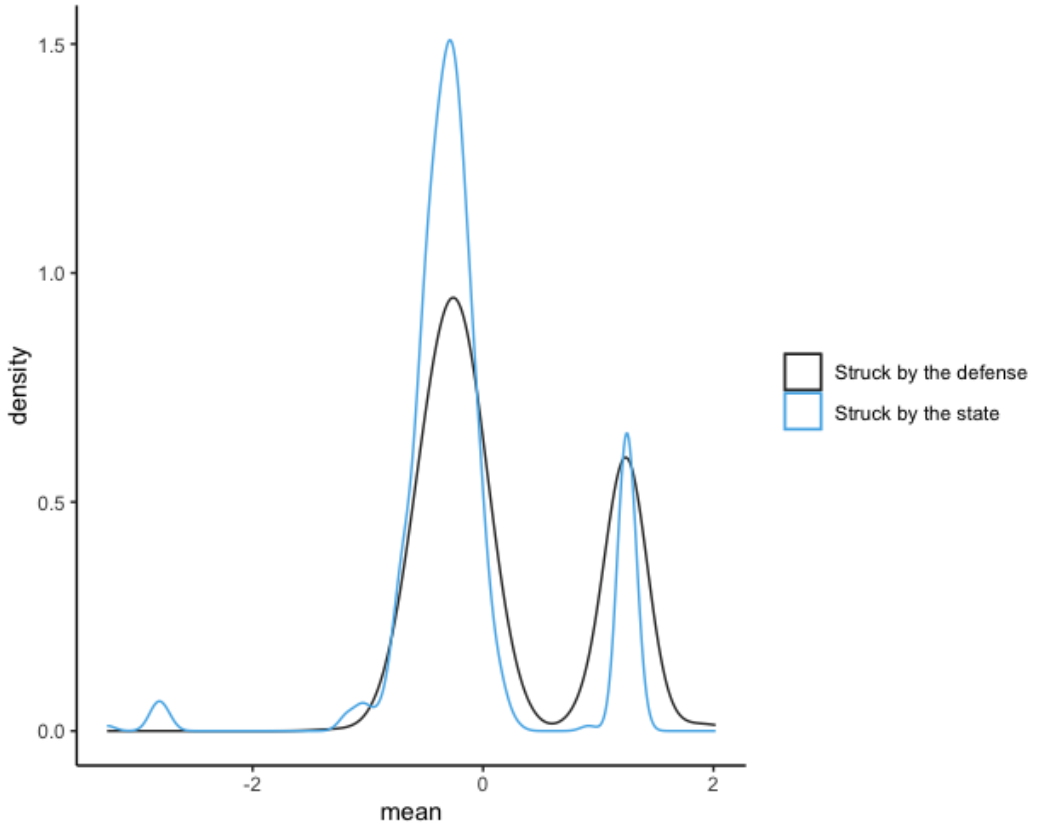


Figure F.1: Density Plot of IRT Proxy by Strikes

[1]IRT analysis was performed using the Political Science Computational Laboratory package: https://cran.r-project.org/package=pscl.

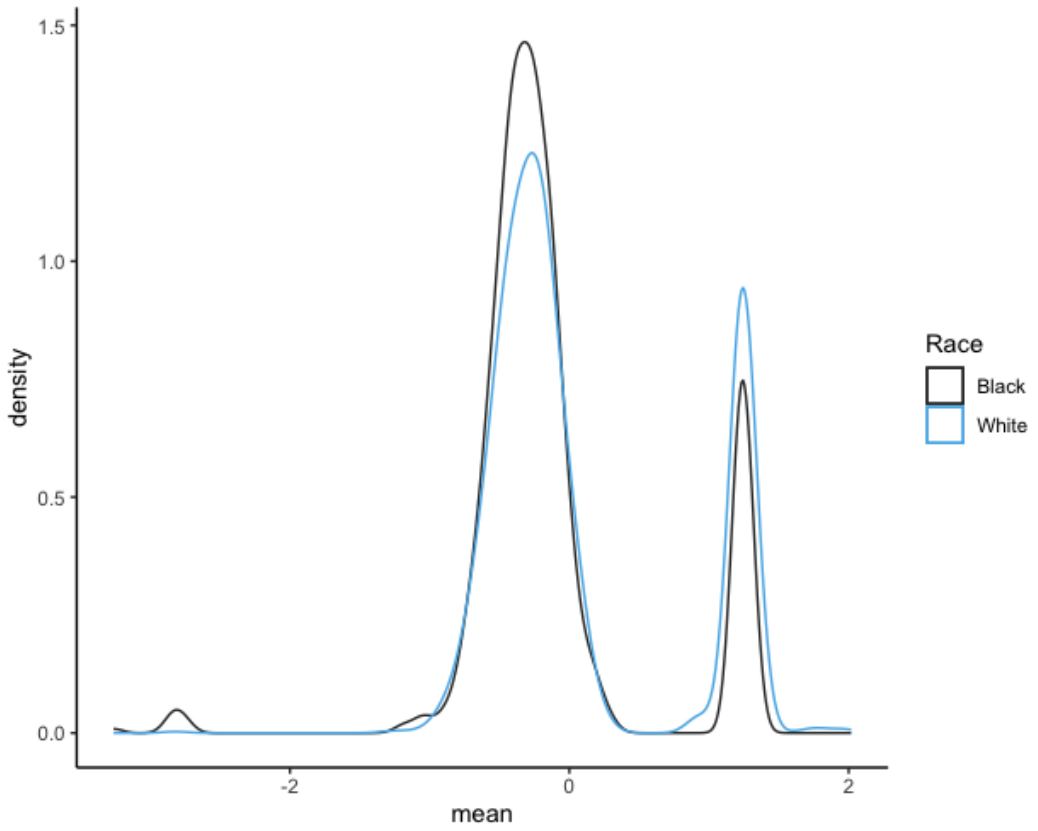


Figure F.2: Density Plot of IRT Proxy by Race

The IRT-based proxy also has a significant pro-defense skew in the selected White jurors and a pro-prosecution—though not significant—skew in selected Black jurors. These skews can be seen in the raw data, as well as in the regression of

$$Proxy_i \sim \beta_0 + \beta_1 White_i + \beta_2 White_i * Selected_i + \beta_3 Black_i * Selected_i$$

The pro-prosecution lean is reduced by 0.142 for selected White jurors as compared to unselected White pool members.

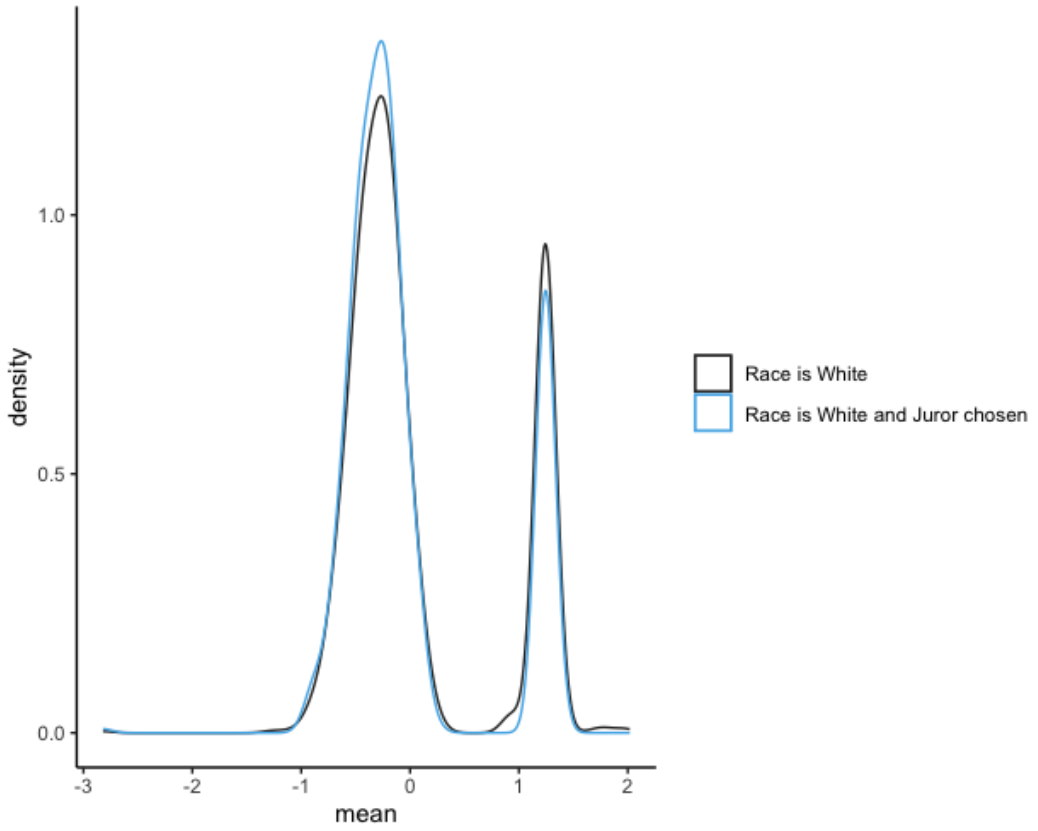


Figure F.3: Plot of IRT Proxy by Selected and Unselected White Jurors

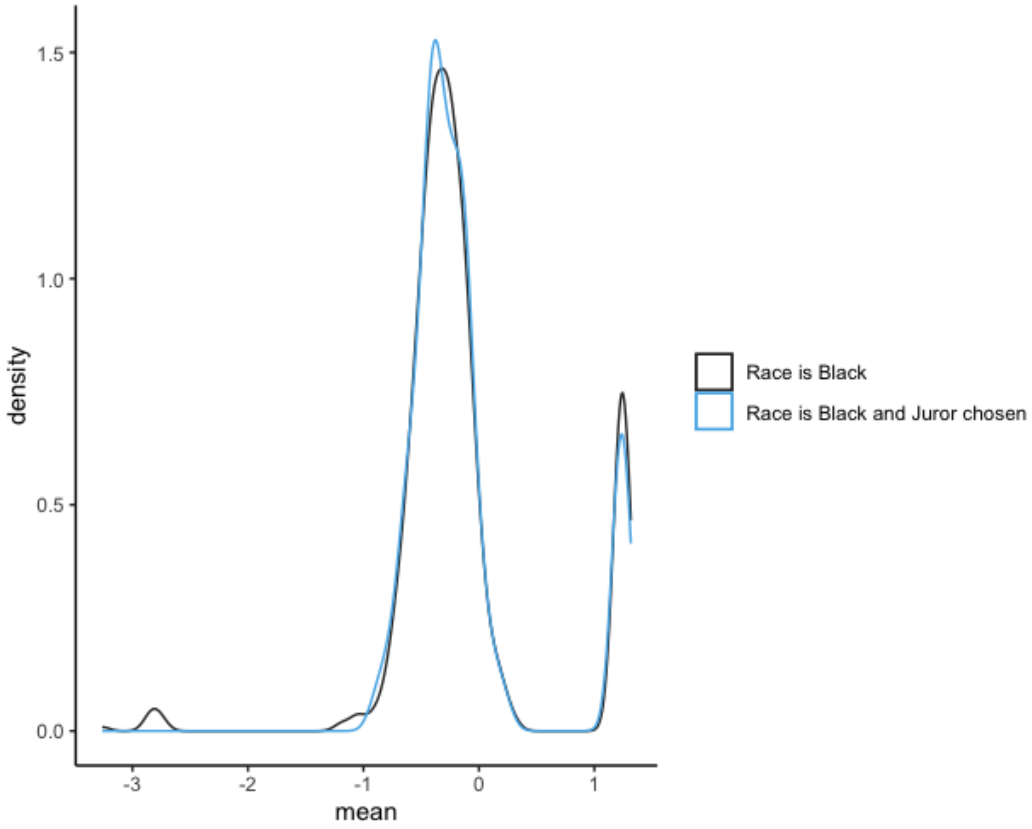


Figure F.4: Plot of IRT Proxy by Selected and Unselected Black Jurors

Table F.1: IRT-Based Proxy Regression Results

	Dependent variable:
	IRT-based Proxy
White	0.267*** (0.040)
White * Selected Juror	−0.142*** (0.035)
Black * Selected Juror	0.030 (0.053)
Constant	−0.143*** (0.033)
Observations	2,279
R^2	0.025
Adjusted R^2	0.023
Residual Std. Error	0.685 (df = 2275)
F Statistic	19.129*** (df = 3; 2275)
Note:	*p<0.1; **p<0.05; ***p<0.01