

The Effects of Panel Assignment on the US Court of Appeals in Death Penalty Cases *

Deborah Beim
Assistant Professor
Department of Political Science
Yale University
deborah.beim@yale.edu

Tom S. Clark
Asa Griggs Candler Professor
Department of Political Science
Emory University
tom.clark@emory.edu

Benjamin E. Lauderdale
Associate Professor
Department of Methodology
London School of Economics
b.e.lauderdale@lse.ac.uk

September 10, 2015

ABSTRACT

PLEASE DO NOT CITE: this paper is based on a preliminary dataset and incomplete analysis.

We use the random assignment of three judge panels on the US Court of Appeals to measure the preferences of individual judges for granting relief in death penalty appeals, and how they are aggregated into decisions. We provide evidence that judges on the US Court of Appeals for the 5th, 6th, 9th, and 11th Circuits apply highly inconsistent thresholds for relief from death penalty sentences. In future versions of this paper, we will examine the extent to which en banc and Supreme Court review reduce the inconsistencies that arise as well as the ultimate effects of random panel assignment on whether and when appellants are executed.

*This version was prepared for the Philadelphia Region American Politics Conference, September 18, 2015.

1. INTRODUCTION

Capital punishment is the most punitive and irreversible form of judicial sanction. Many countries no longer use the death penalty. For those that do, it is clear that its application must meet the very highest standards of fairness and justice. Indeed, it was out of a concern about unfair application of the death penalty that the US Supreme Court, in *Furman v. Georgia* (1972), struck down all death penalty statutes in the country. In a concurring opinion in that case, Justice Stewart wrote, “These death sentences are cruel and unusual in the same way that being struck by lightning is cruel and unusual. I simply conclude that the Eighth and Fourteenth Amendments cannot tolerate the infliction of a sentence of death under legal systems that permit this unique penalty to be so wantonly and so freakishly imposed.” When, just four years later, in *Gregg v. Georgia* (1976), the Supreme Court approved of newly written statutes governing sentencing procedures in 1976, the majority specifically argued that the new statutes addressed the problem of arbitrariness in death penalty sentencing. The crimes for which the death penalty could be applied would be clearly specified and separate sentencing stages for guilt and punishment in capital murder trials would ensure that juries would be able to assess sentencing independently of guilt.

It remains a subject of considerable dispute whether the post-*Furman* reforms in capital sentencing have indeed established a fair, even, and consistent application of the death penalty in the US (Radelet and Pierce 1991; Iyengar 2011; Alesina and Ferrara 2014; Beim and Kastlelec 2014; Canes-Wrone, Clark, and Kelly 2014). In this paper, we examine one aspect of the capital sentence appellate process that has the potential to induce considerable arbitrariness in the application of the death penalty: the extended series of state and federal appeals that always follow an initial conviction and sentence. We focus on the US Court of Appeals, which due to the limited number of appeals heard by the US Supreme Court, is usually the last court to rule on the merits of death penalty appeals before a prisoner

is executed. We show that the evidence from cases heard since *Gregg* imply that different judges apply different standards. On each circuit, there are judges who are more “liberal”, and panels with more of these judges grant relief more frequently; there are also judges who are more “conservative”, and panels with more of these judges grant relief less frequently. The more liberal judges tend to have been appointed by Democratic presidents, the more conservative judges by Republican presidents. Because of the effectively random manner in which death penalty cases are assigned to judges at the courts of appeals, this level of appellate review creates considerable variation in the standards that are applied at this critical stage in the post-conviction review of death sentences. What is more, this the effect of panel composition is not small: in the least consistent circuit we calculate that 29% of death penalty appeals would have been decided differently by the circuit sitting en banc than they were by the panels that heard those cases.

The consequence is that, independent of whether the standards applied by liberal or conservative judges is most faithful to the law, the US Court of Appeals grant relief in death penalty cases in an inconsistent manner. However, that fact alone does not show that executions themselves are administered inconsistently. After a three judge panel submits its decision, the entire Circuit on which they sit can elect to hear the case en banc and the US Supreme Court can choose whether to grant cert in these cases. In principle, one or both of these oversight mechanisms could have the effect of regulating the appeals decisions so as to compensate for the inconsistency created by random panel assignment. In future versions of this paper, we will assess this hypothesis, considering whether panel composition and dissents issued in the courts of appeals are able to provide a sufficient signal to the Supreme Court about which cases to take and potentially reverse in order to achieve consistent standards. In those future versions of this paper, we will also be able assess the extent to which random panel assignment in the Courts of Appeals has causal effects on when and whether prisoners are executed.

2. FEDERAL APPEALS IN DEATH PENALTY CASES

A primary consideration underlying contemporary American death penalty jurisprudence is consistency. This is the central criterion applied in *Furman* and *Gregg*, the foundational contemporary Supreme Court death penalty decisions. The Batson study, famous for showing that defendants who kill white victims are more likely to be sentenced to death, is primarily focused on whether sentences are consistent (Baldus, Woodworth, and Pulaski 1985). Consistency across judges and fidelity to the law—despite variation in preferences and attitudes—is also seminal in the study of judicial politics, especially in the context of the death penalty (George and Epstein 1992).

A number of factors threaten to induce randomness and inconsistency into the application of the death penalty at the sentencing stage, including variation in public opinion (Brace and Boyea 2008; Canes-Wrone, Clark, and Kelly 2014) and state laws (Nice 1992). Only someone who commits a capital murder can be sentenced to be executed, and although there are many federal requirements that reduce variation in sentencing among capital murders, the requirement that a death sentence must be imposed by a jury induces the potential for jury composition to create inconsistency. Considerable evidence suggests the vagaries of subjectivity among juries can increase capricious sentencing. Defendants who kill black victims are less likely to be sentenced to death than defendants who kill white victims (Baldus, Woodworth, and Pulaski 1985; Alesina and Ferrara 2014). When there is a black member of a jury pool, the defendant in question is less likely to be convicted than when the jury pool is all white (Anwar, Bayer, and Hjalmarsson 2012). As public opinion on the death penalty shifts, juries may become more or less likely to rely on it (Baumgartner, Boef, and Boydston 2008).

While jurors are responsible for recommending a death sentence, judges oversee the process and can overrule a jury's recommendation at the sentencing phase itself or afterward

on appeal, typically lowering the sentence to life in prison. In some states, a judge could even override a jury's recommendation of life in prison in favor of imposing a death sentence (see e.g. Radelet 1985; Bright and Kennan 1995, and cites *infra*). In theory, this could either decrease or increase inconsistency, depending on whether juries or judges are more variable in the standards that they apply. Following sentencing, the federal appellate process—especially the habeas corpus process following a death sentence in a state court—is intended, in part, to bring these sentences into line by maintaining standards for procedural fairness and ensuring that minors and the the mentally incompetent are not executed.

Because of these doctrines, nearly all death penalty cases enter a long appeals process after the initial conviction and sentencing phase. Gelman et al. (2004) show that variation early in the death penalty process leads to frequently successful at the appellate levels as well as a lengthy period of review. Of the 6,000 death sentences set down between 1973 and 1995, only 5% had been executed by 1995. Most of the remaining 95% were either overturned or under continued appellate review. This appellate process is often characterized by claims that are very unlikely to prevail, and as a consequence Congress passed, in 1996, the Antiterrorism and Effective Death Penalty Act (AEDPA), which contained procedural hurdles including a one-year statute of limitations period for seeking habeas corpus and severely restricted the ability to file a second or subsequent petition for a writ of habeas corpus, as well as making it more difficult to meet the standards for a writ.

Generally speaking, death penalty habeas corpus petitions reach the courts of appeals after the convicted defendant has exhausted all state-level appeals. As noted, the court of appeals decisions are in nearly all instances the *de facto* final chance to get judicial relief from the death penalty, and these decisions are almost always made by a panel of three judges randomly selected from the pool of judges in the circuit.¹ Past research shows that, when

¹The US Courts of Appeals are arranged into 12 geographically-defined circuits. Currently sitting judges from the circuit responsible for the state seeking the death penalty are the primary pool for any case; however judges from US District Court, from other US Court of Appeals circuits, or retired judges from the circuit

combined with the possibility of oversight from the Courts of Appeals sitting en banc, or from the Supreme Court, randomly assigned panels dramatically increase consistency in outcomes versus single judge decisions (Beim and Kastellec 2014). However, previous research has not compared panel decisions to the natural alternative of having the full circuit decide all death penalty cases.

The question of how panel decisions compare to en banc review is related to the question of how judges influence one another within panels. Judges within a panel may independently apply their own standards with case outcomes determined by majority vote, or they may influence each others' decision-making through deliberation. (Sunstein et al. 2006) finds that judges do not influence one another's decision-making in death penalty cases; (Fischman 2013) and (Beim and Kastellec 2014) find that they do. The deliberative process of collegial decision making on these courts may help promote a uniform standard across the cases heard in a circuit. However, even such influence is not necessarily sufficient to achieve a uniform standard across a circuit if judges pre-deliberation standards vary widely, because randomly assigned three judge panels will frequently group judges with similar views together. The question of influence among the judges on a panel is intertwined with the question of how we measure which standards individual judges are inclined to apply, given that we only observe them making decisions in panels. Thus, in order to evaluate how random assignment to a panel affects the outcome of an appeal as well as subsequent outcomes such as the likelihood and timing of execution, we also need to re-evaluate the question of *preference aggregation on panels*.

sometimes sit as one of the three judges "by designation".

3. THE DATA

Our data include 2,184 death penalty cases decided in the Courts of Appeals between 1983 and 2012.² Because some states sentence far more people to death than others and circuits are organized geographically, cases are very unevenly distributed across circuits. The D.C. Circuit, and the First and Second circuits (which are in the Northeast) yield nearly no cases in our data. We will focus on four circuits with the largest number of death penalty appeals in our data, 68% of the total. These are the Fifth Circuit (MS, LA, TX), the Sixth Circuit (KY, MI, OH, TN), the Ninth Circuit (AK, AZ, CA, ID, HI, OR, MT, NV, WA) and the Eleventh Circuit (AL, FL, GA). We collected the data by searching Westlaw for death penalty cases following the procedures in (Fischman 2013; Beim and Kastellec 2014), coding the judges on the panel and the panel's decision.

We focus on the panel's decision to grant or deny relief to a death-row prisoner. By analyzing the panel's decision rather than individual judges' votes, we maintain a focus on the substantively interesting object, but we forgo individual-level analyses that can consider concurrences and dissents. One reason to approach the problem this way is that judges may suppress dissent for strategic reasons, making observed dissent an imperfect measure of disagreement with the panel decision. A judge might disagree with a decision but know it has no prospect of being overturned, and thus may reserve explicit dissent for those cases where en banc review or Supreme Court might bring the case outcome in line with their preferred outcome. While we do not use the observed dissents in the analysis, the way in which we assess how preferences of judges are aggregated means that we can assess the probability that individual judges disagree with the panel decision even if they do not dissent, as well as assess panel effects (Cross and Tiller 1998; Beim and Kastellec 2014).

²Data collection is still in process. We will ultimately collect all death penalty cases decided in this time period, as well as subsequent outcomes for the defendants.

Although internal rules vary slightly, cases in the Courts of Appeals are assigned to panels of three judges. In the Third Circuit, for example, assignment occurs as follows: “The clerk will use a computer program to randomly select a panel from a pool of all possible three-judge combinations consisting of circuit judges in active service and those judges who have taken senior status and have indicated their willingness to hear death penalty cases” (*Internal Operating Procedures of the United States Court of Appeals for the Third Circuit*). Not all circuits use computerized randomization (Hall 2010). If cases are assigned intentionally, this could inhibit analysis of judges’ preferences: if some judges get systematically weaker cases, they will appear more conservative by virtue of often denying relief. Available evidence suggests that the pseudo-randomizations are sufficient to prevent problematic imbalances in the strength of cases heard by different judges. Republican appointees and Democratic appointees are equally likely to see defendants who won at the district court level (Beim and Kastellec 2014). We will be able to assess this question of pre-treatment balance further once our data set is complete.

4. MEASURING JUSTICES’ REVEALED PREFERENCES FROM PANEL DECISIONS

As described above, in order to measure individual judges’ preferences from aggregate decisions, we need a model for how individual judges preferences aggregate into those decisions, so that we can infer the former from the latter. We use a 1D case-space framework (Kornhauser 1992), which allows us to theoretically describe the preferences of judges and to map different possible preference aggregation processes onto likelihood estimators for each judge’s preferred threshold θ_i for granting relief (). We set up our spatial model in such a way as to make the numerical value of these θ_i equivalent to the rate at which that judge would grant relief, if they ruled alone, given the distribution of cases in their circuit. Our estimates of individual judges’ preferences will therefore be circuit-specific and we will only

compare consistency within circuits, not across them.

We describe each case j as having *facts* that can be described as a location ψ_j . We treat smaller values of ψ_j as indicating stronger appeals (case facts), and larger values of ψ_j as weaker appeals. Each judge i has *preferences* that can be described as a cutpoint θ_i . Each judge, if deciding the case alone, would rule in favor of the appellant if and only if $\psi_j < \theta_i$. Thus, judges with lower cutpoints θ_i are inclined to grant relief in fewer appeals, and judges with higher cutpoints are inclined to grant relief in more appeals. An assumption of this unidimensional model is that all judges agree on the ranking of relative merits of appeals, and disagree only on the appropriate threshold to apply. It is clearly possible for judges to disagree about the relative merits of cases in addition to the threshold at which an appeal is strong enough to require a grant of relief. We choose this modelling strategy because it yields the most favorable estimates for the consistency of the court, given the observed data. To the extent that judges not only disagree about the legal threshold for relief, but also disagree about the relative merits of different requests for relief, we will underestimate inconsistency.

Let $\tilde{\theta}_j$ be the collective decision rule of the panel in case j . This is itself a threshold in the same space of ψ_j , and is assumed to be a function of the values of θ for the three judges hearing case j , which we will refer to as $\theta_{i1(j)}$, $\theta_{i2(j)}$, and $\theta_{i3(j)}$. That is, the thresholds that the individual judges would apply as to when to grant relief are aggregated in some way in order to generate a threshold that will be applied by the panel. This yields the following likelihood function for the set of decisions to grant relief:

$$\mathcal{L}(\theta) = \prod_j p(\psi_j < \tilde{\theta}_j)^{y_j} \cdot p(\psi_j > \tilde{\theta}_j)^{1-y_j} \quad (1)$$

Without loss of generality, we define the scale of case facts ψ to be standard uniform. This implies that judges' preferences θ are also defined on the unit interval, and can be directly

interpreted as the fraction of cases, within the population of cases the circuit considers, where that judge would grant relief if they made the decision alone. This enables us to simplify the above expression as:

$$\mathcal{L}(\theta) = \prod_j (\tilde{\theta}_j)^{y_j} (1 - \tilde{\theta}_j)^{1-y_j} \quad (2)$$

Aside from the fact that $\tilde{\theta}_j$ is a(n as yet undefined) function of the assigned judges for case j , this is the standard Bernoulli likelihood. The model follows this form because we do not attempt to measure or model exactly which cases are strong and which are weak, we instead rely on the randomization of assignment to guarantee that we can talk about the *relief rate* among the cases that the circuit considers.

The crucial question is how the three judges' values of θ map into the decision-rule $\tilde{\theta}_j$ for the panel. In principle, a range of functions are possible and correspond to more or less coherent ways that the judges' individual views might aggregate into a decision. If the panels applied a unanimity rule for granting relief, the implied function would be $\min(\theta_{i1(j)}, \theta_{i2(j)}, \theta_{i3(j)})$, and the most conservative judge would determine the outcome. If the panels applied a unanimity rule for denying relief, the implied function would be the maximum of the individual judges' preferences, and the most liberal judge would determine the outcome. More plausibly, given the fact that the decisions are ultimately majority rule, the implied function would be $\text{median}(\theta_{i1(j)}, \theta_{i2(j)}, \theta_{i3(j)})$, where the median judge's preferences determine the outcome. However, this is a rule where there is no opportunity for influence or panel composition effects, so we also consider a model where we use the mean of the judges' preferences, on a logistic scale, to capture the idea that there might be influence among the panelists rather than a straightforward vote on the basis of ex ante fixed decision rules.³

³The minimum, maximum, and median models are scale-invariant, but the mean model is not, and there are many ways it could be defined. We define the mean on the logistic scale to avoid floor and ceiling effects.

Because we have set up the model this way, we can calculate a statistic describing the extent to which random assignment of judges leads to inconsistency in the decisions that the court makes. The panels' threshold $\tilde{\theta}_j$ for each case is the fitted value for that case: the probability of a grant of relief knowing the panel composition, but nothing about the case facts. If these are all the same, then all panels grant relief at the same rate regardless of composition. If these vary, that is a sign of inconsistency. We cannot know from the data which cases are being decided differently than they would be if the court were consistent, but we can nonetheless calculate what fraction of cases are being decided differently in the aggregate. We define $\bar{\theta}_j$ as the threshold that would be applied by a consistent court, which we define to be the decision-rule that the circuit would follow if deciding cases en banc under the same decision-rule. This leads to an *inconsistency rate* that can be calculated as the mean absolute deviation of the panel thresholds from the en banc threshold:

$$\mathcal{E} = \frac{1}{M} \sum_{j=1}^M |\tilde{\theta}_j - \bar{\theta}| \quad (3)$$

We adopt two approaches to estimating these models. We use maximum likelihood estimation (MLE) because it allows us to formally test the null hypothesis that all judges apply the same threshold for deciding cases and that any variation in case outcomes is the result of random variation in which justices happen to get stronger or weaker cases in sample.⁴ It also allows us to use standard model comparison tools like AIC to assess whether the minimum, maximum, median or mean models seem to best fit the process by which judges' preferences are aggregated into a panel decision.

However, while maximum likelihood estimation facilitates these tests, the individual preference estimates of judges under MLE tend to overstate the degree of variation in the thresh-

⁴We have verified that the likelihood ratio tests have the reported Type I error rate for our model by running models on a large number of placebo data sets created under the null hypothesis by randomly reassigning the relief status across the set of cases.

olds that judges apply. This is a well-known problem with “fixed effects” estimation of large numbers of parameters. Each individual judge’s threshold is estimated with some error, and those errors tend to make the judges appear more varied than they actually are. Therefore, once we have formally tested (and, as we will see, rejected) the null hypothesis that there is no variation in judges’ decision-making rules within circuits, for the purposes of estimating the degree of disagreement and the implied level of inconsistency we use a Bayesian hierarchical model. We model the distribution of judges’ thresholds for granting relief as normally distributed on a log-odds scale, within each circuit. This corrects for spurious variation in judges’ estimated thresholds that results from estimation uncertainty by shrinking the variation in the estimated thresholds towards the mean judge to the extent to which those thresholds are in fact uncertain.

5. VALIDATING RANDOM ASSIGNMENT

Our estimation strategy relies on the assumption that cases are assigned to judicial panels effectively randomly. It is not a problem if certain panels of judges are formed more or less often because of internal administration of the court. However it is a problem if, in expectation, certain judges or panels receive systematically stronger or weaker appeals than others. If this were to occur, the judges receiving stronger cases would grant relief more frequently, the judges receiving weaker cases would grant relief less frequently, and it would appear that they were applying a different legal threshold even though they were not. Past research has shown no difference in the rate at which Democratic and Republican appointees hear appeals that were successful at the District Court level. In future versions of this paper we will assess whether there is any evidence of this kind of problem by examining if there are any observable differences between the cases that are received by different panels before they hear the case.

Because this assumption of random assignment is so critical to the validity of our results, we report the key test here rather than after our main findings. The clearest test we can provide is to use exactly the same methods we use below, but instead of employing the decision of the US Court of Appeals panel as the outcome y_j , we instead use the decision of the US District Court case that led to the appeal. This is a good test because a strong indicator of whether the Appeals court is likely grant relief is whether the District court did. If certain judges are receiving more cases that were decided for the inmate and were appealed by the state, rather than decided for the state and appealed by the inmate, that would be a major concern for our subsequent analysis.

[To be completed in a future version of the paper]

6. MODEL SELECTION

For all four circuits we examine, we can decisively reject the null hypothesis that there is no variation in the grant relief rate. Table 1 shows the p-values for Likelihood Ratio Tests of the minimum, maximum, median and mean models, for each circuit. All tests are highly significant.⁵ For example, if we assume the median judges' preferences determine case outcomes, the p-values for the null of no variation in grant relief rate range from a high of $p = 0.002$ in the 5th Circuit to a low of $p = 0.0000000001$ in the 6th Circuit. The identities of the judges on the panels clearly predicts the probability that relief is granted which—given random assignment—implies that there are causal effects of panel assignment on the outcomes of individual cases.

We cannot apply Likelihood Ratio Tests of the different models against one another, as

⁵This is even stronger evidence than it first appears, because all four circuits have a relatively large number of judges who are sitting by designation from other federal courts. We must estimate a parameter for each of these, which does little to improve the fit of the model due to the rare appearance of that judge, but which increases increases the degrees of freedom of the test just as much as the judges actually pointed to the circuit.

	5th AIC	Circuit p-value	6th AIC	Circuit p-value	9th AIC	Circuit p-value	11th AIC	Circuit p-value
Minimum	548	6×10^{-4}	280	5×10^{-5}	381	4×10^{-5}	558	5×10^{-8}
Maximum	559	8×10^{-3}	272	5×10^{-6}	403	4×10^{-3}	575	6×10^{-6}
Median	551	2×10^{-3}	234	1×10^{-11}	380	3×10^{-5}	552	8×10^{-9}
Mean	548	6×10^{-4}	250	4×10^{-9}	360	2×10^{-7}	543	4×10^{-10}

Table 1: Comparison of Preference Aggregation Models. AIC for models fit by MLE on each circuit, based on assuming that the median, mean, minimum, maximum preferences of the three judges on the panel determine the outcome. Likelihood ratio test p-values compare each model to a null model in which the decisions of all panels on the circuit apply the same threshold regardless of which judges sit on the panel.

they are non-nested, however when we compare them by AIC we observe that the best fitting models are not identical across circuits. On the 5th, 9th, and 11th Circuit, the mean model fits best, while on the 6th Circuit, the median model fits best. Generally, the minimum and maximum models fit far worse, which is not surprising as they is no particular reason to expect the panels to follow a unanimity rule.⁶

We can think of the difference between the median aggregation rule and the mean aggregation rule as the difference between a model with no panel effects (median) and a particular model of panel effects (mean). The median model implies that the non-median judges’ views only matter through the determination of which judge is the median. Thus the mean model reflects more ‘consensual’ panels, while the median implies a more ‘adversarial’ panel. This is broadly consistent with past scholarship that has shown a lack of panel effects on the 6th Circuit relative to other circuits—for example, (Epstein, Landes, and Posner 2011) show that the dissent rate is highest in the Sixth Circuit. These weak panel effects have been attributed to a relatively high level of personal animosity among the judges on that circuit (Lane 2001). However, we should emphasize that a direct statistical comparison of the median and mean

⁶For the 5th Circuit, the minimum model fits as well as the mean model, but for all other circuits it fits far less well than either the median or mean model. The 5th Circuit shows the lowest level of variation in thresholds across judges and therefore the variation in fit across model specifications is lower for that circuit than for the others.

models is not unequivocal for any of these comparisons.⁷

We will proceed with using the mean model for the 5th, 9th, and 11th circuits, and the median model for the 6th. We note however that the analysis that follows is not sensitive to this decision: the differences between these two models of how panels aggregate preferences among the three judges of a panel have negligible effects on our estimates of the inconsistency rate of the court. For most cases, the median preferences and the mean preferences of the judges are very similar, it is only in cases where there are a pair of judges very close together and a third judge far to the left or right of them that these models make distinct predictions.

7. STATIC ESTIMATES OF JUDGES' REVEALED PREFERENCES

In Figure 2 we show the estimated preferences of 5th, 6th, 9th, and 11th Circuit judges who sat on at least 10 panels in our data set. These are colored by the party of the appointing president. The 6th Circuit shows a strikingly strong partisan division, and a truly massive gap in judges' individual preferences. On that court, there are four Democrat appointed judges who (when the median) grant relief in over 90% of cases, while there are also nine Republican appointed judges who (when the median) grant relief in less than 15% of cases. None of the other Circuit show either such a high level of disagreement or disagreement that is as strongly associated with the appointing presidents' parties. However, all the other Circuits show a level of variation in thresholds that is substantial from the perspective of legal consistency, with preferred grant rates varying from below 10% on the low end to above 60% on the 5th Circuit, 80% on the 11th Circuit, and 90% on the 9th Circuit.

These preference estimates are conditional on the fitted models and the assumed aggre-

⁷If we use a Vuong test for non-nested models to assess the relative fit of the two models, the probabilities of observing the improvements in fit that we see for the mean model over the median model, assuming that both are actually equally close to the true data generating process, are 0.40, 0.94, 0.07, and 0.17 for the 5th, 6th, 9th, and 11th Circuits respectively.

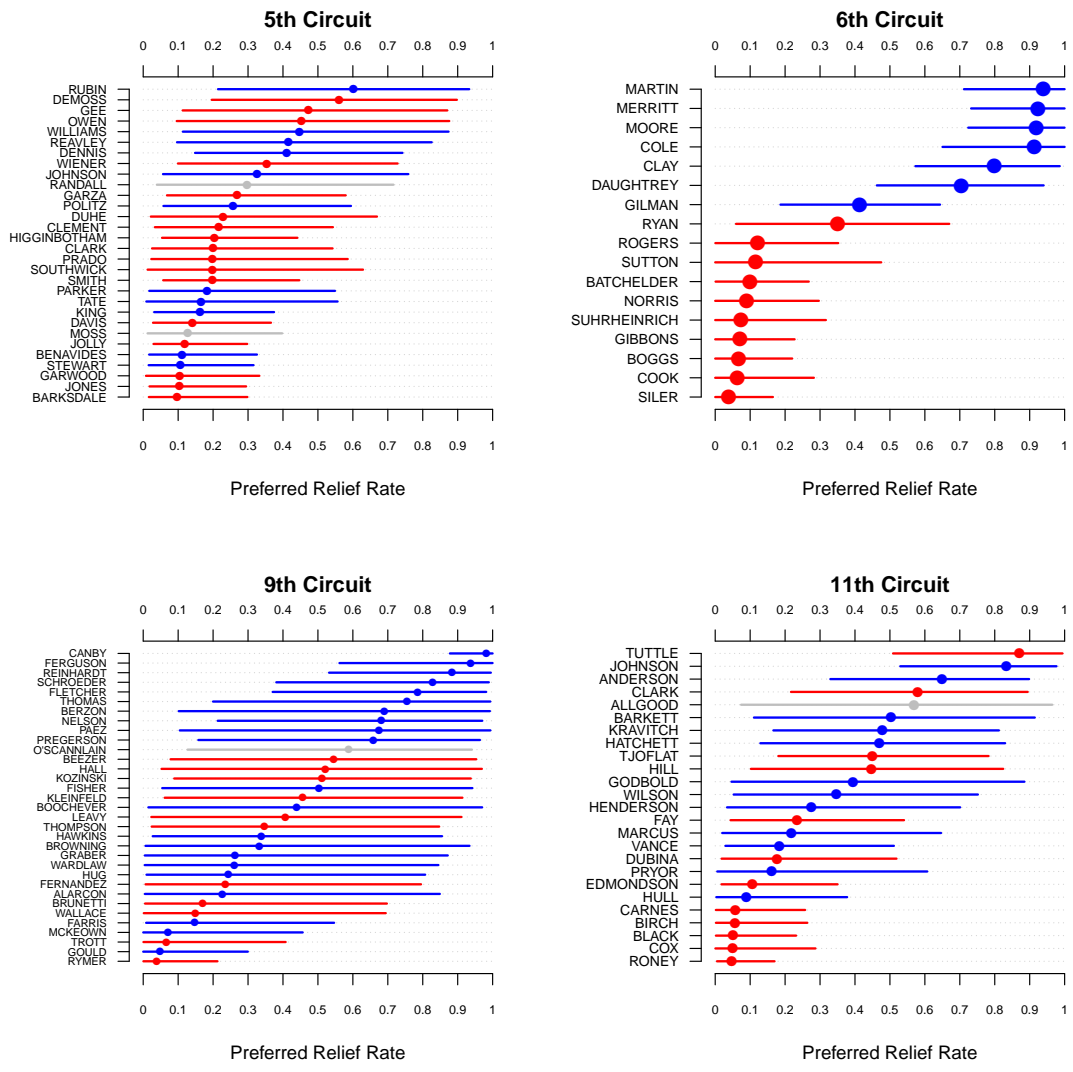


Figure 1: Preferred relief rates for judges sitting on at least 10 death penalty appeals panels.

gation rules. Thus, to the extent the aggregation rules are approximations, the estimates will be as well. However, even though this is true, this cannot constitute an argument against the claim that there is variation in the thresholds imposed by different panels. These are the judges' preferences that, given the assumed preference aggregation rule, would yield the observed patterns of decisions. If panel composition had no predictive effect, all these estimates would collapse to their common mean.

8. DYNAMIC ESTIMATES OF JUDGES' REVEALED PREFERENCES

One possible way in which static estimates of judge's preferred relief rates could be biased is if the mixture of case facts has changed over time such that appeals became systematically stronger or weaker. If this were the case, judges who served during only part of the historical period we examine would have faced a different distribution of cases, leading to different estimated preferred rates of relief, even though they applied the same threshold. To make sure this is not driving the results above, here we explore a variation on the estimation approach which allows judges preferred grant rate to vary over time. The motivation here is not to assess whether the judges' legal thresholds are changing in an absolute sense, but to assess whether there are systematic trends that are attributable to the appeals they receive becoming stronger or weaker. This also enables us to account for legal changes like the passage of AEDPA, which changed the legal thresholds for the appeals we examine. This analysis ensures that we are primarily assessing the relationship between panel composition and decisions across cases decided at about the same time, rather than potentially relying on inter-temporal comparisons that might reflect legal changes or historical trends in the composition of the death row population.

[To be completed in a future version of the paper]

9. ESTIMATES OF PREDICTED PROBABILITY OF RELIEF BY PANEL COMPOSITION

In Figure 2, we show the distribution of predicted probabilities for receiving relief, across the observed panels in our data set. These plots demonstrate the consequences of the variation in judges' preferred relief rates for the prospects of appeals. Again, the 6th Circuit is the most extreme case. Because there is a large cluster of (Republican appointed) judges who prefer to grant relief very rarely, there is a substantial mode of cases facing panels with at least two such judges for which the median judge's preferences, and thus the predicted probability of relief, is less than 20%. However, because there is a smaller cluster of (Democrat appointed) judges who prefer to grant relief in almost all cases, there is also a substantial mode of cases facing panels with at least two such judges with very good prospects of relief.

It is only by comparison to the 6th Circuit that the other Circuits could be said to be consistent. Depending purely on which judges are on the panel for a given case, the prospects for appeals vary substantially in all Circuits. The lowest inconsistency rate—the fraction of cases that were actually decided differently than they would have been by an en banc decision of the Circuit—is 8% on the 5th Circuit. This is a non-trivial level of randomness introduced by judges' disagreements about legal thresholds. The 16%, 22%, and 29% inconsistency rates on the 11th, 9th, and 6th Circuits respectively are substantial by any plausible standard for the consistency of legal decision-making. Indeed, the maximum theoretically possible inconsistency rate as we have defined it is 50%, on a circuit exactly balanced between judges/panels who always prefer to grant relief and those who never prefer to grant relief. This analysis reveals—very starkly—that there is substantial and consequential disagreement among judges on the US Court of Appeals as to the threshold that individuals facing the death penalty must meet in their appeals. Because these cases are randomly assigned, the strength of the cases faced by each panel and by each judge is the same in

expectation. The implication of these results is that different panels apply widely varying legal thresholds.

While these three judge panels on the US Court of Appeals are effectively the court of last resort for many of these appeals, some are heard by all judges on the Circuit sitting en banc, and some are granted cert to be heard by the US Supreme Court. So even though the inconsistency rates demonstrated above are very high, they are not the final outcome of the legal process. In the following sections, we examine the extent to which the inconsistency introduced by the panels is resolved by further oversight, before turning to the effects of panel assignment on the probability and timing of execution. Before proceeding though, we can already see that these inconsistency rates imply a very high burden for en banc and Supreme Court review. If the observed rates of such review are lower than the rates at which panels are making decisions inconsistent with what the circuit would decide en banc, it is not possible for these oversight mechanisms to fully regularize these panel decisions. Even if the observed rates of review are higher, the appropriate cases must be identified to review. In the following two sections, we examine two kinds of available signals that might be used, before turning to the long-run consequences of panel assignment on the probability and timing of execution.

10. DISSENT AS A SIGNAL FOR EN BANC OR SUPREME COURT OVERSIGHT

One way that the inconsistencies observed above might be reduced is if having a dissent on a panel triggers oversight, either by the entire Circuit sitting en banc, or by the US Supreme Court.

[To be completed in a future version of the paper]

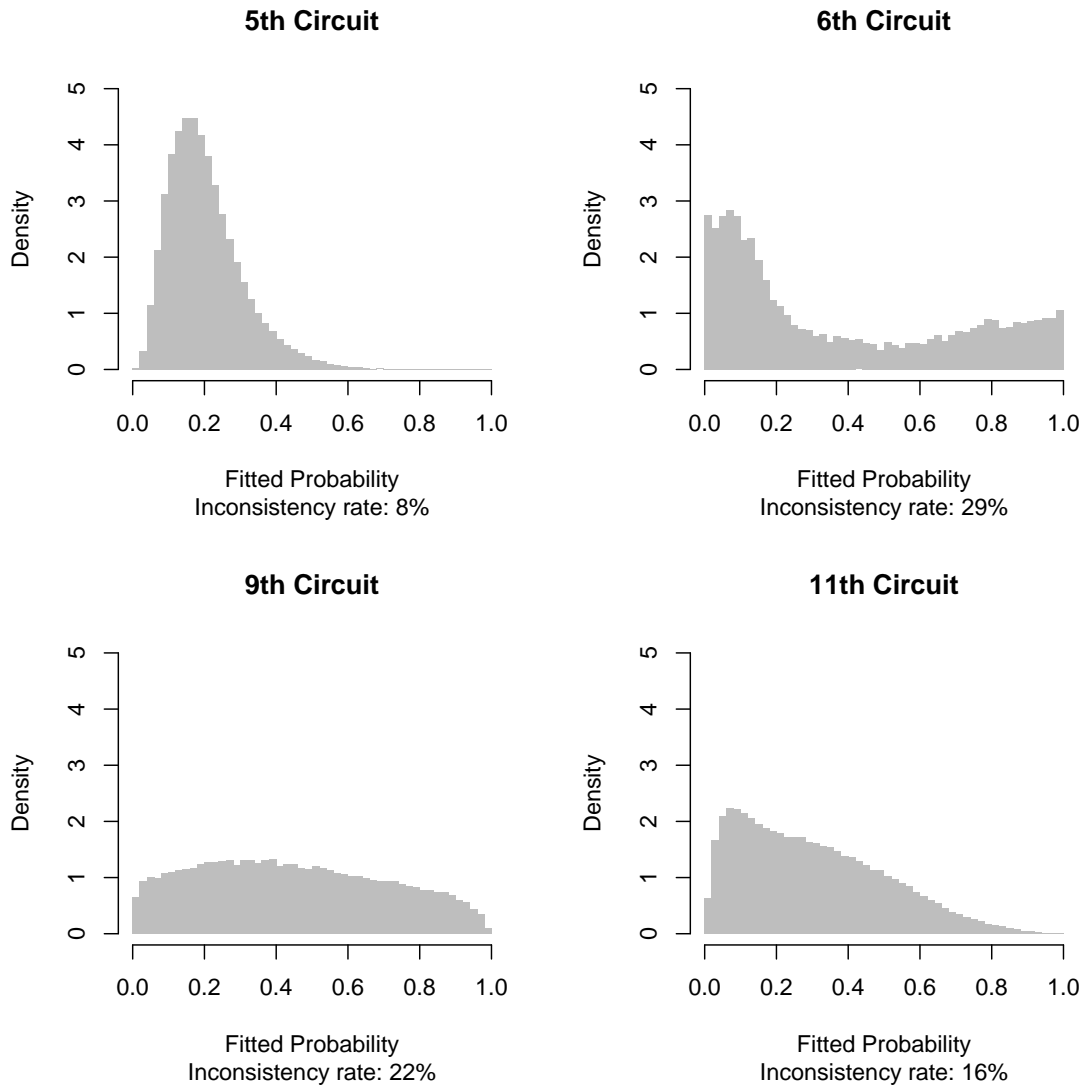


Figure 2: Predicted probabilities of receiving relief, conditional on panel composition.

11. PANEL COMPOSITION AS A SIGNAL FOR EN BANC OR SUPREME COURT OVERSIGHT

Dissent is limited as a means of signalling the need for oversight in a given case because it requires having someone on the panel inclined to send the signal. If all three judges on a panel are more liberal or all three are more conservative than the consensus of the circuit, they will decide some cases differently than the en banc circuit would without any judge dissenting. The same logic applies relative to the Supreme Court. In such situations though, the composition of the panel could itself trigger oversight by the entire Circuit sitting en banc, or by the US Supreme Court.

[To be completed in a future version of the paper]

12. PANEL ASSIGNMENT EFFECTS ON PROBABILITY AND TIMING OF EXECUTION

The ultimate question at stake in these appeals is whether and when the appellant will be executed. Here we examine the consequences of panel assignment on probability and timing of execution, measured from the panel decision. Some cases reach the panel in a final appeal to stay an imminent execution while others are less urgent, however we can once again rely on random assignment to guarantee that there is no systematic relationship between the average proximity to execution and the composition of a panel before the panel hears the case.

[To be completed in a future version of the paper]

13. CONCLUSIONS

[To be completed in a future version of the paper]

References

- Alesina, Alberto and Eliana La Ferrara (2014). “A Test of Racial Bias in Capital Sentencing”. *American Economic Review* 104.11, pp. 3397–3433.
- Anwar, Shamena, Patrick Bayer, and Randi Hjalmarrsson (2012). “The Impact of Jury Race in Criminal Trials”. *Quarterly Journal of Economics* 127, pp. 1017–1055.
- Baldus, David C., George Woodworth, and Jr. Pulaski Charles A. (1985). “Monitoring and Evaluating Contemporary Death Sentencing Systems: Lessons from Georgia”. *UC Davis Law Review* 18.1375.
- Baumgartner, Frank R., Suzanna L. De Boef, and Amber E. Boydston (2008). *The Decline of the Death Penalty and the Discovery of Innocence*. Cambridge University Press.
- Beim, Deborah and Jonathan P. Kastellec (2014). “The Interplay of Ideological Diversity, Dissents, and Discretionary Review in the Judicial Hierarchy: Evidence from Death Penalty Cases”. *Journal of Politics* 76.4.
- Brace, Paul and Brent D. Boyea (2008). “State Public Opinion, the Death Penalty, and the Practice of Electing Judges”. *American Journal of Political Science* 52.2.
- Bright, Stephen B. and Patrick J. Kennan (1995). “Judges and the Politics of Death: Deciding between the Bill of Rights and the next election in capital cases”. *BUL Rev* 75.759.
- Canes-Wrone, Brandice, Tom S. Clark, and Jason P. Kelly (2014). “Judicial Selection and Death Penalty Decisions”. *American Political Science Review* 108.1, pp. 23–39.

- Cross, Frank B. and Emerson H. Tiller (1998). “Judicial Partisanship and Obedience to Legal Doctrine: Whistleblowing on the Federal Courts of Appeals”. *Yale Law Journal* 107, pp. 2155–76.
- Epstein, Lee, William M. Landes, and Richard A. Posner (2011). “Why (and When) Judges Dissent: A Theoretical and Empirical Analysis”. *Journal of Legal Analysis* 3.1, pp. 101–137.
- Fischman, Joshua B. (2013). “Interpreting Circuit Court Voting Patterns: A ‘Social Interactions’ Framework”. *Journal of Law, Economics, & Organization*.
- Gelman, Andrew et al. (2004). “A Broken System: The Persistent Patterns of Reversals of Death Sentences in the United States”. *Journal of Empirical Legal Studies* 1.2, pp. 209–261.
- George, Tracey E. and Lee Epstein (1992). “On the Nature of Supreme Court Decision Making”. *American Political Science Review* 86.2, pp. 323–337.
- Hall, Matthew (2010). “Randomness Reconsidered”. *Journal of Empirical Legal Studies* 7.3, pp. 574–589.
- Internal Operating Procedures of the United States Court of Appeals for the Third Circuit*. Effective 2015. URL: http://www2.ca3.uscourts.gov/legacyfiles/2015_IOPs.pdf.
- Iyengar, Radha (2011). “Who’s the Fairest in the Land? Analysis of Judge and Jury Death Penalty Decisions”. *Journal of Law and Economics* 54.3, pp. 693–722.
- Kornhauser, L. A. (1992). “Modeling Collegial Courts II: Legal Doctrine”. *Journal of Law Economics & Organization* 8, pp. 441–70.
- Lane, Charles (2001). “Disorder in the Court”. *Washington Post*.
- Nice, David C. (1992). “The States and the Death Penalty”. *Western Political Quarterly* 45.4, pp. 1037–1048.
- Radelet, Michael L. (1985). “Rejecting the Jury: The Imposition of the Death Penalty in Florida”. *UC Davis Law Review* 18, p. 1409.

- Radelet, Michael L. and Glenn L. Pierce (1991). "Choosing Who Will Die: Race and the Death Penalty in Florida". *Florida Law Review* 43.1, pp. 1–34.
- Sunstein, Cass R. et al. (2006). *Are Judges Political? An Empirical Analysis of the Federal Judiciary*. Washington, D.C.: Brookings Institution Press.