How Consistently Are Death Penalty Appeals Decided?

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

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

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

July 28, 2018

ABSTRACT

We use the quasi-random assignment of cases to three-judge panels on the US Courts of Appeals to measure variation in the thresholds that individual judges apply when granting relief in death penalty appeals. We find clear evidence that panels vary in the thresholds they apply for granting relief in these cases as a function of which judges are assigned to the panel. These inconsistencies in panel decisions are substantial. We further find strong evidence that the ultimate probability of execution depends on the judges that are assigned to an inmate’s first Court of Appeals case in at least some Circuit courts.
1. INTRODUCTION

Capital punishment is the most punitive and irreversible form of judicial sanction. Therefore, 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, the majority specifically argued that the new statutes addressed the problem of arbitrariness in death penalty sentencing.

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 Kastellec 2014; Canes-Wrone, Clark, and Kelly 2014). In this paper, we examine one aspect of the appellate process that has the potential to either increase or decrease arbitrariness in the application of the death penalty. We focus on the US Court of Appeals, which, due to the limited number of cases heard by the US Supreme Court, is usually the court of last resort for those who face execution.

Consistent with a generation of research showing that judges exhibit ideological variation in their voting patterns generally (Segal and Spaeth 2002), we show that different judges apply different standards to death penalty cases. 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 initially assigned to panels at the Court of Appeals, this level of appellate review creates variation in the standards that are applied at this critical stage in the post-conviction review of death sentences. The effect of panel composition is not small: a substantial fraction of death penalty appeals would have been decided differently by a hypothetical average judge (e.g. the median of all judges on the circuit) 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 Courts of Appeals grant relief in death penalty cases in an inconsistent manner.

We go further and estimate the consequences of this heterogeneity on whether individuals are executed. The existence of variation in the standards judges apply to death penalty appeals alone does not necessarily 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 (Beim and Kastellec 2014). Bearing in mind the complexity of these institutions that mitigate inconsistency, we focus on the reduced-form results for the prisoners themselves and assess the extent to which random panel assignment in the Courts of Appeals has a causal effect on whether prisoners are executed. While the effects for execution are somewhat smaller than for relief due to these oversight mechanisms and variation in state procedures, we nonetheless find evidence that panel composition predicts execution probability for cases initially reaching the Court Appeals in the same year, in at least some circuits, via multiple estimation strategies. This finding is the first to our
knowledge to systematically document the consequences of panel composition for the way in which criminal justice is administered to individual defendants. This finding also calls into question the extent to which the American legal system is meeting the standard that the Supreme Court has set out for the death penalty’s conformity to the Eight Amendment’s prohibition on cruel and unusual punishment.

2. FEDERAL APPEALS IN DEATH PENALTY CASES

After the Supreme Court’s 1976 decision in in *Gregg*, and its companion cases validating a handful of new death penalty statutes, the number of criminal defendants in the US sentenced to death grew steadily until the late-1990s and has been decreasing since. Similarly, the number of executions taking place grew during the last quarter of the twentieth century, peaking at 98 in 1999, and has been declining since. Baumgartner et al. (2018) document a number of striking patterns in the use of the death penalty in the United States during this period. They show that these trends in death sentences hold despite a relatively constant rate of defendants’ sentences being overturned. Further, the amount of time that elapses from sentencing until an execution is carried out has increased steadily over time. In recent years, those executed have spent, on average, approximately 20 years on death row. Finally, and perhaps most troubling, Baumgartner et al. (2018) find a great deal of variation in the number of death row inmates exonerated across the states, suggesting a great deal of variation in states’ internal conviction and sentencing procedures. The high rates of sentences being overturned, the long procedural delays, and variation in exonerations all call into question whether the application of the death penalty meets the standards laid out by the Court in *Gregg*.

As Baumgartner et al. (2018) argue, the spirit of the Court’s decisions validating new death penalty statutes in 1976 included the proposition that the imposition of the death
penalty cannot be arbitrary or random. 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). 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 Boydstun 2008).

Gelman et al. (2004) show that variation early in the death penalty process leads to frequent success in subsequent appeals 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. There is substantial documented variation in the likelihood of execution conditional on a death sentence, which varies especially strongly by state (see e.g. Spurr 2002; Smith 2012). 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 successive petition for a writ of habeas corpus, as well as making it more difficult to meet the standards for a writ.¹

In this paper, we aim to understand a different potential source of variation in the pro-

¹One consequence of the enactment of the AEDPA for our analysis is that it could potentially shorten the time from a defendant’s first appearance before the Court of Appeals and his or her ultimate execution. The modeling strategy we employ below accounts for any resulting change in systematic temporal patterns around 1996.
cess: variation in judicial behavior on the Courts of Appeals. Generally speaking, death penalty cases reach the Court of Appeals after the convicted defendant has exhausted all state-level appeals. As noted, the Courts of Appeals decisions are almost always 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 selected from the pool of judges in the circuit.\(^2\) These courts have the opportunity to correct for error and inconsistency that has occurred previously, but also potentially to introduce inconsistency themselves. Previous work has demonstrated that the existence of panels reduces inconsistency in death penalty cases versus a hypothetical alternative of single-judge decisions in the Courts of Appeals (Beim and Kastellec 2014). That paper also demonstrates that en banc and Supreme Court review help reduce inconsistency by preferentially reviewing decisions by panels that are likely to have made extreme decisions. However, the remaining degree of inconsistency due to variation in judges’ decision-making is important, and so in this paper, we aim to quantify the inconsistency that remains by comparison to a hypothetical of consistent decision-making by all judges within a circuit. In doing so, we recover novel (albeit imprecise) estimates of the rates at which individual judges would prefer to grant relief, if they decided cases individually. Further, we are able to estimate the causal effect of the arbitrary assignment of judicial panels on the probability of execution for those appealing their death sentences.

The question of how panel decisions differ is, of course, 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 in more complex ways through deliberation. Sunstein et al. (2006) argues that judges do not influence one another’s decision-making in death penalty

---

\(^2\)The 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 circuits of the U.S. Courts of Appeals, or retired judges from the circuit sometimes sit as one of the three judges “by designation”.

5
cases; Fischman (2015) and Beim and Kastellec (2014) argue 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 consider the question of preference aggregation on panels. Although we have an understanding of the consequences of individual ideology on criminal justice outcomes (see e.g. Huber and Gordon 2004; Cohen and Yang 2018) and on defendants’ long-run outcomes (see e.g. Aizer and Doyle 2015); and a notion of panel effects influence how a case is ultimately resolved (Hall 2009), we do not have a firm understanding of the reduced-form, ultimate consequences of panel composition.

3. ESTIMATING THE CAUSAL EFFECT OF PANEL ASSIGNMENT ON INMATE FATE

We estimate the consequences of panel composition on decisions and executions by comparing cases that were assigned to different judges. Although internal rules vary, cases in the Courts of Appeals are assigned to panels of three judges randomly or quasi-randomly. In the Third Circuit, for example, assignment occurs as follows: “The clerk will use a computer program

---

3Our work is related to a larger literature on panel effects on the Courts of Appeals, see e.g. Cross and Tiller (1998), Kastellec (2007), Kim (2008), and Boyd, Epstein, and Martin (2010), including papers that consider solutions for the inconsistency generated by panel assignment (Hasday 2000; Revesz 2000).
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) or publicise their procedures. Levy (2017) writes that judges are randomly assigned in all death penalty cases in the Second, Third, Fourth, and Eleventh Circuits. In the Fifth, Sixth, Seventh, Ninth, and Tenth Circuits, judges are randomly assigned but the assignment also takes into account concerns such as workload and the variety of judges with which an individual sits each year. It is not a problem for our identification strategy if panels themselves are not randomly constructed, so long as cases are assigned to those panels without respect to the facts of those cases. The most serious threat to our inferential strategy would be if cases are assigned in a way that causes some judges to get systematically weaker cases than other judges, because those judges would spuriously appear more conservative by virtue of more frequently denying relief. Available evidence suggests that the pseudo-randomizations are sufficient to prevent problematic imbalances in the strength of cases heard by different judges. In the domain of death penalty cases, previous research has indicated that Republican appointees and Democratic appointees are equally likely to see defendants who won at the district court level (Beim and Kastellec 2014).

We exclusively examine the first appearance of a given death-row inmate before the Courts of Appeals. Some inmates have multiple cases heard over a period of years, but whether further cases are heard for the same inmate can be an outcome of the initial case. Further, panel assignment is not independent across these cases and the strength of such cases are unlikely to be independent either.4 By focusing on the initial cases only, we are able to adopt the most credible feasible identification strategy for the causal effect of the

4This is especially true for cases decided after the passage of AEDPA, for reasons discussed above.
judges assigned to hear an appeal on the subsequent outcomes of interest.

Thus, the identification strategy we follow leverages the assumption that assignment of cases to panels is as-if-random, at the moment of initial panel assignment. This means that, among a set of cases considered at the same time by the same appellate court, the expected strength of the cases heard by any possible three-judge panel is the same. This statement does not hold across circuits or over time: the mix of appellants is different in different parts of the US and changes over time. Further, the Supreme Court precedents that the Appeals courts apply are also changing over time. This means that we must analyse each Circuit separately and we must also adjust for the average “case strength” at any given moment in time, in order to isolate any effect of which judges are assigned to the panel for a given case. In this section, we describe how we build a model for appeals court decisions from assumptions about the decision rules followed by the judges and about the evolution of case strengths over time.

3.1. Data

To evaluate the causal effect of panel assignment on inmate fate, we constructed a new database of death penalty cases in the United States.\(^5\) We performed an over-inclusive Westlaw search for all cases that could be a case from a death row inmate before any circuit in the US Court of Appeals following the procedures in Fischman (2015) and Beim and Kastellec (2014).\(^6\) This procedure yielded more than 20,000 cases between 1983 and 2012. Each case was then read, individually, by either one of us or a member of our research team that comprised law students, graduate students in political science, and undergraduate

\(^5\)There are many databases of death penalty appeals in existence. However, as we describe below, we collect several pieces of information that is not available in other databases.

\(^6\)Following Fischman’s procedure (and Beim and Kastellec) we searched on all Courts of Appeals (the CTA database in Westlaw): (CAPITAL /S (PUNISHMENT MURDER CRIME OFFENSE)) (SENTENC/ /S DEATH) & da(aft 1/1/1983 & bef 12/31/2012).
students. Each case was assessed for whether it was a death penalty case. We retained each case decided by a three-judge panel (i.e., we exclude cases decided *en banc*) and recorded a number of pieces of information, which we describe below. We then further examined all cases identified by a similar search decided between July 2, 1976 (when the Supreme Court decided *Gregg v. Georgia* 428 U.S. 153) and 12/31/1982, to see if any of our death penalty defendants had a case in the Court of Appeals prior to 1983. If they did, we excluded them from our data. As a result, our data includes those individuals who appeared with death penalty-related cases for the first time before the courts of appeal between 1 January 1983 and 31 December 2012.\(^7\)

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, Second and Third circuits (covering the mid-Atlantic through New England) yield too few cases for us to study. Therefore, we focus on the eight circuits—the Fourth through Eleventh, inclusive—which handle nearly all death penalty cases. The resulting data include 1991 death penalty cases decided in the Court of Appeals between 1983 and 2012. Table 1 reports the number of defendants that come from each state, along with the number of petitioners for whom relief was denied in their initial appellate case and the number in which the petitioner was ultimately executed. Slightly more than half of the petitioners who appear in our data set (1011) were ultimately executed.

For each case, we recorded the names of the judges participating in the decision, how each judge voted (whether to support any relief at all for the defendant), and whether the panel majority decided to support any relief at all for the defendant.\(^8\) We also recorded the}\(^7\)Our data exclude all cases in which a person was sentenced to death if that sentence was commuted prior to the appeal. Our data also exclude cases brought by next friends and cases in which the person on death row does not seek relief (such as cases about prison conditions).

\(^8\)Note, each judge could individually support relief on *some* grounds, but the majority might still fail to agree on any single grounds for relief, a phenomenon known as the doctrinal paradox (Kornhauser and Sager 1986).
<table>
<thead>
<tr>
<th>Circuit</th>
<th>State</th>
<th># Petitioners</th>
<th>Relief Denied</th>
<th>Petitioner Executed</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>MD</td>
<td>10</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>NC</td>
<td>90</td>
<td>79</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>SC</td>
<td>33</td>
<td>29</td>
<td>24</td>
</tr>
<tr>
<td>4</td>
<td>VA</td>
<td>111</td>
<td>99</td>
<td>91</td>
</tr>
<tr>
<td>5</td>
<td>LA</td>
<td>46</td>
<td>32</td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>MS</td>
<td>42</td>
<td>27</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>TX</td>
<td>512</td>
<td>431</td>
<td>356</td>
</tr>
<tr>
<td>6</td>
<td>KY</td>
<td>17</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>MI</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>OH</td>
<td>118</td>
<td>73</td>
<td>44</td>
</tr>
<tr>
<td>6</td>
<td>TN</td>
<td>32</td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>IL</td>
<td>53</td>
<td>47</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>IN</td>
<td>25</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>AR</td>
<td>50</td>
<td>35</td>
<td>22</td>
</tr>
<tr>
<td>8</td>
<td>IA</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>MO</td>
<td>119</td>
<td>97</td>
<td>71</td>
</tr>
<tr>
<td>8</td>
<td>ND</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>NE</td>
<td>13</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>SD</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>AK</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>AZ</td>
<td>52</td>
<td>30</td>
<td>26</td>
</tr>
<tr>
<td>9</td>
<td>CA</td>
<td>66</td>
<td>37</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>ID</td>
<td>6</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>MT</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>NV</td>
<td>9</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>OR</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>WA</td>
<td>10</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>CO</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>KS</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>NM</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>OK</td>
<td>145</td>
<td>107</td>
<td>92</td>
</tr>
<tr>
<td>10</td>
<td>UT</td>
<td>6</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>WY</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>AL</td>
<td>75</td>
<td>61</td>
<td>46</td>
</tr>
<tr>
<td>11</td>
<td>FL</td>
<td>212</td>
<td>157</td>
<td>68</td>
</tr>
<tr>
<td>11</td>
<td>GA</td>
<td>115</td>
<td>67</td>
<td>46</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td>1991</td>
<td>1512</td>
<td>1011</td>
</tr>
</tbody>
</table>

Table 1: Distribution of cases and outcomes by circuit and state.
name of the defendant in the case, the state of origin, procedural history (which other courts heard the case), the date of the panel’s decision, who the district (federal trial) court judge was, and how s/he voted. One crucial piece of information is the name of the inmate in the case. The most exhaustive databases of death penalty appeals, such as the one maintained by the Bureau of Justice Statistics, do not include the name of the inmate at hand in the case. (They only include identification numbers which cannot be linked to identities or court records.) Our collection of that information allows us to study the fate of individual defendants.

We focus on two outcomes. The first outcome is the panel’s immediate 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 of 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.9

The second outcome of interest is whether the death-row prisoner is ultimately executed. This outcome allows us to test whether the quasi-random assignment of judges matters, once the full judicial review process has been exhausted. We employ the list of executions maintained by the Death Penalty Information Center (https://deathpenaltyinfo.org/) current as of 2017, which we verified by cross-referencing against states’ websites listing executions. We then matched each execution to an inmate appearing in our data. We only

9While 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 could 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).
use the fact of the inmate’s execution in the analysis below, but these data also include other information like the date of execution, the inmate’s age, and the method of execution.

The baseline rates at which inmates are granted relief and the rates at which they are executed vary over the period we study within and across circuits. Figure 1 shows these trends for each circuit separately. Note that towards the end of the data series, the execution rates decline in several circuits. This may be a lasting trend, or it may reflect the fact that some of those cases will result in execution, but have not done so yet. These plots highlight the importance of not assuming comparability of cases heard at different points in time: as-if-random assignment of cases to panels only ensures comparability of cases reaching the same Court of Appeals for the first time at the same moment in time.

Inmates who were granted relief are not necessarily removed from death row. The relief granted might be, for example, a remand to reconsider whether a defense attorney’s representation was competent. Nevertheless, whether an inmate is granted relief in their initial appeals court case and whether they are ultimately executed are highly correlated outcomes. Table 2 reports the proportion of inmates ultimately executed in each circuit (as of 2017) divided into those who were granted relief at their first appeal before the Court of Appeals and those who were not. In all but one circuit, the initial panel decision is associated with a 20-33 percentage point difference in execution rates, while the 10th Circuit has a much stronger association between having the appeals court deny relief and execution (56 percentage points). The result of the first appeal before the Court of Appeals is strongly associated with execution.

Our question, though, is not whether there is variation in execution rates as a function of how a panel decides a case. Presumably, if case outcomes are driven by the merits of an appeal, then we should expect lower rates of execution among those appeals that are successful. Instead, our question is if the identity of the judges to which a death row inmate is randomly assigned has a causal effect on the inmate’s immediate and ultimate fate.
Figure 1: Five-year moving average for the rate at which death row inmates are denied relief by the Courts of Appeals, and at which they are ultimately executed, by year of initial Courts of Appeals case decision.
In order to link individual judges’ preferences to decisions made by panels of those judges, we need a model for how individual judges’ preferences aggregate. This is a particularly vexing problem at the US courts of appeals, where there is a well-documented pattern of dissent suppression. That is, judges might disagree, but for a variety of potential reasons might choose not to record their dissenting votes. As a result, judges’ votes may generate a misleading picture of disagreement if used in models that aim to recover judges’ preferences. This problem has plagued scholars seeking to measure the preferences of judges at the courts of appeals (see e.g. Fischman 2015).

Competing theoretical models of judging at the Courts of Appeals propose different incentive structures for judges that implicate how they aggregate their preferences when deciding a case. For example, strategic models of the judicial hierarchy suggest the potential for Supreme Court oversight creates incentives for judges to vote against their sincere preferences under certain circumstances (Songer, Segal, and Cameron 1994; Cameron, Segal, and Songer 2000; Carrubba and Clark 2012). The incentive to adhere to the preferences of Supreme Court justices, as opposed to an appellate judge’s own sincere view, could lead to

<table>
<thead>
<tr>
<th>Circuit</th>
<th>Execution Rate</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Granted Relief</td>
<td>Denied Relief</td>
</tr>
<tr>
<td>4th</td>
<td>33%</td>
<td>64%</td>
</tr>
<tr>
<td>5th</td>
<td>40%</td>
<td>72%</td>
</tr>
<tr>
<td>6th</td>
<td>10%</td>
<td>43%</td>
</tr>
<tr>
<td>7th</td>
<td>0%</td>
<td>29%</td>
</tr>
<tr>
<td>8th</td>
<td>28%</td>
<td>58%</td>
</tr>
<tr>
<td>9th</td>
<td>17%</td>
<td>37%</td>
</tr>
<tr>
<td>10th</td>
<td>20%</td>
<td>76%</td>
</tr>
<tr>
<td>11th</td>
<td>20%</td>
<td>48%</td>
</tr>
</tbody>
</table>

Table 2: Proportion of inmates executed in each circuit, according to whether they are granted any relief at the initial appeal.

3.2. Modeling Panel Decisions
“whistleblowing” by a dissenting judge at the Court of Appeals (Beim, Hirsch, and Kastellec 2014; Beim and Kastellec 2014). The Supreme Court is not the only overseer to which appellate judges might be attentive. All active judges in the circuit could decide to collectively review the panel’s decision *en banc* (Clark 2009; Beim, Hirsch, and Kastellec 2016); as such, the full circuit serves as another principal. Other arguments, though, suggest that appellate judges might be more myopic, not responsive to incentives created by the hierarchical organization of the courts (Hettinger, Lindquist, and Martinek 2004). The consequence of these various claims is that it is unclear whose preferences the panel decision represents. The decision could be governed by median of the panel, but it might also represent the preferences of an extreme judge—who could be a whistleblower—or the decision might be some other function of the judges’ individual preferences that is the outcome of some bargaining and preference aggregation process. Our inability to know how the judges’ votes aggregate poses challenges for how we can draw inferences about individual preferences from collective outcomes.

Our approach avoids most of the problem of insincere voting by judges and allows us to explore a range of possible aggregation rules that might relate judges’ preferences to their collective decision. We use a one-dimensional 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 for granting relief (Hangartner, Lauderdale, and Spirig 2015). The key innovation is to define a latent dimension substantively, such that positions to one end are those that indicate a greater propensity to grant relief, and then to take advantage of the shuffling of judges into different panels across cases.

Formally, we describe each case $j$ as having *facts* that can be described as a location $\psi_j$. We treat smaller values of $\psi_j$ as indicating stronger appeals (case facts), and larger values of $\psi_j$ as weaker appeals. Each judge $i$ has *preferences* that can be described as a cutpoint $\theta_i$. 
Each judge, if deciding the case alone, would rule against the inmate if and only if \( \psi_j < \theta_i \). Thus, judges with lower cutpoints \( \theta_i \) are inclined to grant relief in more appeals, and judges with higher cutpoints are inclined to grant relief in fewer 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 modeling 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 \( \bar{\theta}_j \) be the collective decision rule of the panel in case \( j \). This is itself a threshold in the same space of \( \psi_j \), and is assumed to be a function of the values of \( \theta \) 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 a single decision \( y_j = 1 \) to deny relief:

\[
\mathcal{L}(y_j | \theta, \psi) = p(\psi_j < \bar{\theta}_j)^{y_j} \cdot p(\psi_j > \bar{\theta}_j)^{1-y_j}
\]  

(1)

How do the three judges’ values of \( \theta \) map into the decision-rule \( \bar{\theta}_j \) 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 \( \max(\theta_{i1(j)}, \theta_{i2(j)}, \theta_{i3(j)}) \), 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 minimum of the individual judges’ preferences, and the most liberal judge would determine the outcome. More plausi-
bly, given the fact that the decisions are formally 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. The median is the rule we adopt for the presented analysis. This is a rule where there are no explicit influence or panel composition effects, however moving to a model that captures such effects (e.g., a mean preference model, or other variations) has negligible effect on any of the quantities that we aim to estimate in this paper.

3.3. Modeling Temporal Change

We have already argued that neither the strength of cases arriving at the Court of Appeals nor the review processes after the Court of Appeals takes its initial decision are likely to be static over time.\(^{10}\) However, the Court of Appeals does not hear enough death penalty cases for us to only look at data from very brief time windows where it would be reasonable to assume that the distribution of case facts $\psi$ is a static distribution. Instead, we assume that the distribution of the case strengths varies smoothly over time: at any given point in time the distribution of case strengths is normal, with mean $\phi_t$ and variance 1. This generates a probit model with a dynamic intercept plus a term that depends on the median preference of the three judges on the panel. Thus, the baseline strength of cases can get weaker or stronger over time, and the effect of the panel composition is variation around that baseline.

We assume that judges’ relative preferences for granting relief are static over their careers, as there are not enough cases to estimate dynamic preferences for individual judges. As with our other assumptions above, this can only bias the estimates away from finding variation in judges’ preferences: if judges’ relative preferences for granting relief are actually changing, they are necessarily inconsistent with one another.

\(^{10}\)For example, as we pointed out, the enactment of the AEDPA in 1996 might have had the consequence of shortening the time from a defendant’s first appearance before the Court of Appeals until his or her execution.
\[ p(y_j = 1) = \Phi \left( \phi_t + \theta_{j}^{median} \right) \]  
\[ \theta_{j}^{median} = \text{med} \left( \theta_{i1(j)}, \theta_{i2(j)}, \theta_{i3(j)} \right) \]
\[ \phi_t \sim N \left( \phi_{t-1}, \sigma_{\phi} \right) \]

Given this model, variation in \( \theta_{j}^{median} \) due to the composition of panels directly maps to a non-zero causal effect of panel assignment on the likelihood an inmate receives any relief from the Court of Appeals.

3.4. Regularization

There is substantial variation in the number of cases heard by different judges over their careers. While there are a few judges who have heard over 100 death penalty appeals, there are also a number of district court judges sitting on a panel by temporary assignment who appear only once in the data set. We cannot precisely estimate the preferences of judges who appear only once or twice, and even those who appear dozens of times will be imprecisely estimated. Given these data constraints, we take two approaches to modeling variation in judges’ preferences: a “judge” model and a “party” model. Both of these put additional structure on the variation in judges’ preferences.

In the judge model, we assume all judges’ preferences come from a common hierarchical model for \( \theta_{i} \):
\[ \theta_{i} \sim N(0, \sigma_{\theta}) \]

Thus, under the judge model, we estimate both a general variation parameter \( \sigma_{\theta} \) and the preference \( \theta_{i} \) of each judge. This is relatively demanding of the data, but enables us to examine individual judges for evidence of variation. The special case \( \sigma_{\theta} = 0 \) corresponds to
the case where all judges have identical preferences for granting relief, and there are thus no effects of panel assignment.

In the party model, we estimate a single parameter $\delta_{RD}$ as the difference in preferences between judges appointed by Republican and Democratic presidents on a circuit, assuming no within-party variation. The special case $\delta_{RD} = 0$ corresponds to the case where all judges have identical preferences and there are no effects of panel assignment. There are two major advantages of this approach versus estimation judge-specific preferences. First, it is a more parsimonious specification and we have more statistical power to detect differences by party of the appointing president. Second, the Republican-Democratic difference can be positive or negative, and thus can be estimated in an unbiased way, whereas the variance parameter $\sigma_\theta$ cannot be negative, and thus it is effectively impossible to generate an unbiased estimator for it in precisely the case we are most interested in: $\sigma_\theta = 0$. The disadvantage of the party model is that we will fail to discover variation across judges that is not a function of appointing presidents’ parties, variation that is just as consequential for the consistency of the appellate process. If there is variation in judges’ thresholds, it will surely not be entirely a function of party of the appointing president. There is good reason to believe such variation exists: not all judges appointed by one party’s presidents have identical views.

We also estimate a third “null” model where all judges are assumed to have the same preferences ($\theta_i = 0$) for purposes of testing whether allowing variation in judges preferences improves model fit. The out-of-sample predictive power of the party and judge models relative to this null model is central to adjudicating our research question: does knowing the identities of the judges on a panel help predict whether they will grant relief or whether an inmate will be ultimately executed?
3.5. *Modeling the Risk of Execution*

The above model describes the decision-making process at the level of the Court of Appeals. What we want to assess, however, is not only the effect of panel composition on receiving relief from the Court of Appeals, but also on the ultimate outcome of whether the inmate is executed. There are multiple ways we could do this. For clarity of argument, we simply substitute the execution outcome \( z_j \) in the place of the panel decision outcome \( y_j \) in exactly the same model described above. This “reduced form” approach has the advantage of directly assessing the ultimate effect of panel composition on the execution outcome. That effect is likely to run partly, but not entirely, through the immediate decision of the panel on whether to grant relief. This is not the only channel because the same panel is more likely to hear subsequent appeals from that defendant and the identities of the judges may influence the subsequent oversight process via en banc and/or Supreme Court review. It is the very complexity of this subsequent process that makes the reduced form analysis we adopt here attractive: it allows us to estimate whether there is a causal effect of initial panel assignment on execution, regardless of the relative importance of the various causal pathways through which that effect could arise.

In doing the analysis this way, the structural logic of the case space model described above no longer applies because the judges on the panel do not make a direct decision about execution. Nonetheless, this approach facilitates tests of whether panel composition has a causal effect on the probability of execution that are comparable to the tests for the panel decision, which is difficult to achieve using models that attempt to describe both outcomes at the same time, or that use preference estimates based on the decision of whether to grant relief to predict execution.
3.6. *Estimation and Hypothesis Testing*

We estimate our models by Bayesian posterior simulation. Estimation by maximum likelihood or maximum posterior is unattractive for this problem because the median preference model does not yield an entirely well-behaved likelihood function, and numerical maximization is consequently difficult. We estimate three models (judge, party and null) for each of the six circuits, for each outcome (relief denial and execution). The models are estimated using Stan (Carpenter et al. 2016), based on MCMC simulations of 2,500 iterations, from which the first 500 are discarded.

As noted above, for the judge model, the parameter value of $\sigma_\theta = 0$, which is equivalent to the null model of all judges applying the same threshold, is at the boundary of the parameter space for a variance parameter. As a consequence, $\sigma_\theta$ is necessarily estimated with upward bias if the true value of the parameter is 0. This is a problem because we would like to clearly establish whether there is evidence against this null hypothesis of consistent adjudication. Therefore, in order to assess whether we have significant evidence of variation in judges’ thresholds under the judge model, we apply a version of randomization inference. We simulate new data sets from the null model (including time trends, but no variation in judges’ preferences) for each circuit, for both relief denial and execution. For each simulated data set, we then estimate the judge model, storing the resulting estimates of $\sigma_\theta$. We can then compare the estimate of $\sigma_\theta$ from the model applied to the real data to the distribution of estimates of $\sigma_\theta$ resulting from data sets we know to have been generated from the null model where all judges apply the same standards and variation in outcomes is purely a function of the observed temporal patterns and random variation in the strength of cases. This allows us to calculate a p-value for how unlikely it is that we would obtain a $\sigma_\theta$ estimate as large or larger than the one that we did, if the null hypothesis $\sigma_\theta = 0$ were true, which addresses any concerns about estimator bias for the $\sigma_\theta = 0$ case. For both outcomes, we simulate
1000 data sets for each circuit, and do the full Bayesian posterior analysis. The circuits can be examined for evidence against consistent decision adjudication either individually or collectively by using Fisher’s combined probability test.

4. RESULTS

4.1. Differences between Democratic and Republican Appointed Judges

We now consider the causal effect of the partisanship of the judges to whom a case is assigned at the Court of Appeals. Figure 2 shows the estimated effect of a change from a Democratic appointed panel majority to a Republican appointed panel majority in each of the circuits we consider. The plots show posterior means and central 50% and 95% intervals. For each circuit, we show the estimated effect on (a) whether the inmate is granted relief at the Court of Appeals, and (b) whether the inmate is ultimately executed.

All but one of the point estimates are positive, indicating a general tendency for Republican-majority panels to be associated with both higher relief denial rates and higher probability of ultimate execution. The estimates are not individually significant for most circuits, but the p-values associated with observing 8 positive coefficients in 8 circuits for relief denial and 7 in 8 for execution are 0.004 and 0.03, respectively. The Sixth Circuit stands out as the one circuit where we find strong evidence of a causal effect of the partisanship of the panel on both the likelihood an inmate is denied relief and the likelihood an inmate is ultimately executed. That is not surprising, others have documented (Yung 2010) that on the Sixth Circuit judges’ partisanship is particularly predictive of their voting, as contrasted with other circuits. In other circuits, we might expect the lines of disagreement to align less neatly with partisanship.

If we weight the circuit-level estimates by the number of cases decided in those circuits,
Figure 2: Estimated treatment effects in each circuit for a change from a Democratic majority to a Republican majority panel, as differences in probability of negative outcome for the inmate.
we can construct an estimate of the average treatment effect of going from a Democratic appointed majority to a Republican appointed majority, for all death penalty cases in our data set. For relief denial, the average treatment effect is 10.9% (95% interval 6.8%-15.4%); for execution, it is 8.5% (95% interval 3.4%-13.2%). These average treatment effects of panel composition as measured in terms of appointing president partisanship are substantial. They correspond to 218 additional relief denials (95%: 136-306) and 168 additional executions (95%: 68-263) in a world where all panels had Republican majorities versus a world where all panels had Democratic majorities.

It is important to reiterate that the individual-level treatment effects underlying these average treatment effects are both heterogenous across circuits and across time, and also that appeals would have faced different probabilities of having panels with Republican or Democratic majorities across circuits and across time. While we have made some functional form assumptions in order to form this estimate, there is no reason to expect those assumptions to bias the estimate away from zero given that the Republican-Democratic difference is symmetrical and could in principle and in estimation be either positive or negative.

4.2. Variation in Judges’ Relative Thresholds for Granting Relief

As we noted previously, a generation of research on measuring the preferences of judges in the US courts has proceeded from the perspective that partisanship, while informative, is at best a rough proxy for judges’ legal and political views. In order to overcome the limitations of studying partisanship, we now turn to our model of individual judges’ thresholds for granting relief in cases from death row inmates.

In Table 3 we show p-values derived from comparing the estimated standard deviation $\sigma_\theta$ of individual judges’ preferences to the randomization distribution for that parameter for data sets generated from the null model. For relief, the p-values are less than 0.1 in
six circuits, less than 0.05 in five circuits, and less than or equal to 0.01 in three circuits. 

Since these are fully independent data sets and analyses, we can use Fisher’s Combined Probability Test to construct a joint p-value for the null hypotheses that all circuits have judges applying identical thresholds for these cases, or equivalently that these p-values are distributed standard uniform. The joint p-value that we calculate is on the order of $10^{-7}$. \(^{11}\)

For the execution outcome, the p-values are less than 0.1 in four circuits and less than 0.05 in one circuit. Still, there is strong evidence collectively across circuits that these p-values do not come from a standard uniform distribution: the joint p-value is 0.01. \(^{12}\)

These results indicate unambiguous evidence against consistent adjudication at the level of panels, and strong evidence these inconsistencies persist in such a way as to induce a causal effect on execution. The evidence is consistent with the evidence from the party model. We now examine some of the evidence with respect to individual judges. In Figure 3 we show the estimated preferences $\theta_i$ from the judge model for the judges who sat on at least 10 panels in our data set. These plots show the relative preferences for granting/denying relief implied by assuming the median of the panel determines the case outcome, holding constant

---

\(^{11}\)This p-value is based on $\chi^2 = 61.6$ with 16 degrees of freedom. For one circuit we never observe an estimate in the randomization distribution that is greater than the estimate on the data, so we set the p-value to half of the inverse of the number of randomizations. This may therefore be an overestimate of the joint p-value.

\(^{12}\) $\chi^2 = 31.9$ with 16 degrees of freedom.
the baseline rate of relief in the year that the case was decided.

The Sixth Circuit shows the clearest association between party of the appointing president and individual judges’ preferences (recall that party is not used in the estimation of the judge model). The judges in this circuit are strongly sorted by party of the appointing president. However, even though party of the appointing president does not sort judges as well in the other circuits, we nonetheless estimate very substantial variation across judges. Because the judge preference estimates are on a standard normal scale, the differences between judges estimated at -1, 0 and +1 would be the differences between granting relief 83%, 50% and 17% of the time, if the preferred grant rate of the median circuit judge were 50%. We see judges estimated at values less than -1 and greater than +1 in many of the circuits. Because of the structure of the data and the limited number of cases, the exact degree of variation in judges’ preferences is not estimated very precisely, but the data are consistent with substantively large variation across judges.

5. CONCLUSION

Because the death penalty entails such a high degree of punitiveness and irreversibility, its exercise requires the highest level of legal scrutiny to ensure its application does not violate individual rights. The Supreme Court has held that the constitutional prohibition on cruel and unusual punishment requires the death penalty not be administered randomly or in a way affected by factors that are orthogonal to the legal merits of the crime and defendant. Our analysis demonstrates two findings that call into question the extent to which the death penalty is administered in a fair and even-handed fashion. First, consistent with scores of research, we find that judges have varying preferences, which correlate with partisanship, and that those preferences are predictive of how judges vote.

We join a literature that shows the causal effect of the identity of the judge on how a
Figure 3: Estimated relative preferences for granting relief among judges sitting on at least 10 death penalty appeals panels. Higher values are higher thresholds for granting relief. Judges appointed to the Courts of Appeals by Republicans are plotted in red, those by Democrats in blue.
panel votes (cf., Boyd, Epstein, and Martin 2010). Our research goes further, to show a causal effect of the identity of the judge for the ultimate fate of the litigants coming before the court. Normative concerns would be lessoned if the identity of the judges were to matter only for how panels decided cases, while the institutions of judicial oversight mitigated any such variation so that individuals ultimately received the same treatment. However, what we find is that being randomly assigned to one judge or another has a causal effect on whether or not an individual lives or dies.

The implications of these findings raise a series of normative concerns surrounding state executions as well as a number of avenues for future research on consistency in adjudication in democracies. We consider three such issues.

**How much inconsistency is tolerable?** We have shown that variation in judges’ preferences results in a causal effect on death row inmates’ legal and personal fates. The luck of the draw in case assignment affects who lives and who dies. Of course, some degree of inconsistency should be expected in the course of any kind of dispute resolution, especially when there are many people who resolve said disputes. What makes this example particularly troubling is the finality of the death penalty and the fact that its exercise is the most extreme use of governmental power against the citizen.

To understand whether the inconsistency we document raises from the level of “troubling” to the level of intolerable, we must define what level of inconsistency Americans are willing to tolerate. The US Supreme Court, for example, has held that capital punishment must not be administered randomly in order to pass constitutional muster. Specifically, it must not be “arbitrary” or “capricious.” (c.f. *Furman v. Georgia* (1972) and *Gregg v. Georgia* (1976).)

To the extent that the luck of the draw from the judicial pool equates with randomness, then any level of inconsistency that is thereby caused is intolerable. There is extensive
scholarship and many U.S. Supreme Court decisions on how consistent the initial sentence must be. But for many people on death row, the length of time waiting for execution induces another source of randomness. Since many of the cases we study here specifically concern the right to a successive habeas petition—a grant of which leads directly to more time on death row—the inconsistency we identify in relief and execution amplifies this inconsistency. The constitutionality of this as-applied randomness has received attention in the federal courts (see for example Jones v. Chappell 31 F.Supp.3d 1050 (C.D. Cal. 2014) and Levy (2015), though see Jones v. Davis 806 F.3d 538 (9th Cir. 2015) reversing that decision) but none from the U.S. Supreme Court. The question of which inmates among those sentenced to death are in fact executed is an under-explored source of inconsistency, both academically and doctrinally. Of course there are alternative normative approaches, beyond constitutional prescription, that one might choose to adopt. That endeavor, though, is beyond the scope of what we can accomplish here. What we can provide is empirical grist for the mill of normative evaluation of the institution.

Moreover, this line of inquiry suggests another empirical approach that could, at a minimum, provide some evidence of the relative severity of inconsistency we document. One might perform an analysis similar to the one we conduct here, but examining another area of law where the stakes are relatively lower—i.e., where a citizen’s corporeal fate is not on the line—and compare whatever causal effect of judge assignment is recovered to the one we document here. To the extent we could perform such analyses on multiple areas of the law, we could hope to make some statement about how bad the degree of inconsistency in the death penalty is relative to inconsistent adjudication in other areas of the law. Of course, as we described above, the data collection efforts necessary for our analysis were substantial, and so this kind of empirical analysis falls far beyond the scope of a single article. One would first need to identify another area of law that generates a similar or greater number of structurally similar cases. In spite of these obstacles, we regard such a line of inquiry as
a valuable effort for future researchers.

**Why can some circuits mitigate the effects of judges on inmate fate?** Our analysis points to another kind of variation beyond comparing how inconsistency in adjudication ranks across different areas of the law. Specifically, our results suggest that in some circuits panel assignment has a larger effect on the legal outcome of the legal case than it does on the inmate’s ultimate fate. What might mitigate the treatment effect from panel assignment in some circuits such that the judges affect the outcome of the petition but less so whether an inmate is executed? In other circuits, though, the magnitudes of these two effects are flipped. What might exacerbate the treatment effect of panel assignment on the inmate’s fate in these circuits?

These questions suggest the need for theoretical analyses of the circuit courts’ internal institutional rules and their effects on within-circuit consistency. Little research has examined how institutional variation within the circuits affects the degree to which circuit courts are able to ensure intracircuit consistency in how they decide cases (for one example, see Strayhorn 2013). Specifically, the various circuits adopt different kinds of procedures governing the extent to which the various judges know about and can express opinions about other panels’ decisions, but we understand little about why the circuits adopt the rules they do and therefore the effects these rules have on how much decisions vary as a function of panel assignment. This lack of investigation is particularly surprising, given long-running concerns about the extent to which the design of the courts of appeals create the kinds of opportunities for the inconsistency we have shown in our analysis (Carrington 1969; Posner 1985).


