Research on the Death Penalty

Chance and the Death Penalty

Richard A. Berk
Robert Weiss
Jack Boger

We address capriciousness in decisions by prosecutors to charge homicide defendants with a capital crime. We suggest that it is useful to think of such decisions as a kind of lottery in which one should focus on the *distribution of outcomes* when considering both the nature of capriciousness and the degree of capriciousness. After our conceptional framework is introduced, we illustrate our ideas with the analysis of a data set from San Francisco.

The death penalty is serious business. It is, of course, very serious for individuals convicted of capital crimes, but it is also taken very seriously by the criminal justice system. Rulings by the Supreme Court of the United States beginning decades ago have asserted that arbitrariness will not be tolerated in the processes by which some individuals are sentenced to death and others are not (*Furman v. Georgia* 1972; *Gregg v. Georgia* 1976; *Woodson v. North Carolina* 1976; *Proffit v. Florida* 1976; *Godfrey v. Georgia* 1980; *Maynard v. Cartwright* 1988). Two kinds of arbitrariness have been identified: the use of inappropriate factors, such as the offender's race, and a capriciousness in which

Earlier versions of this article were given at the U.C.L.A. Marschak Colloquium, at the Neyman Seminar, University of California, Berkeley, Department of Statistics, and the August 1990 meetings of the American Statistical Association. We very much appreciate the thoughtful comments that were made at each. We are especially indebted to William Mason, David Freedman, and John Rolfe for a number of technical suggestions and to Anthony Amsterdam for help in our legal interpretations. The data were made available through the efforts of Michael Burt and Grace Suarez, from the Office of the Public Defender, San Francisco. Alec Campbell did much of the initial file construction. Weiss's work was supported by a grant from the U.S. Department of Health and Human Services (grant MH37188-06).

Direct correspondence to Richard A. Berk, Center for the Study of the Enviroment and Society, University of California, Los Angeles, Haines Hall 264, 405 Hilgard Avenue, Los Angeles, CA 90024.

Law & Society Review, Volume 27, Number 1 (1993)

offenders are sentenced to death in an unsystematic fashion. In both instances, there is no "principled" way to distinguish those who are sentenced to death from those who are not, although the first is unprincipled because of the use of unacceptable determining factors, while the second is unprincipled because of an absence of any manifest rationale.

The vehicles by which the Supreme Court has addressed the issue of potential arbitrariness have been legal challenges to particular death penalty statutes (Gross & Mauro 1989:ch. 1). In *Furman v. Georgia*, all existing statutes were implicitly ruled unconstitutional because they failed to constrain the discretion of juries, thereby allowing those juries to mete out sentencing decisions that were too often "cruel and unusual" in their result; they were an invitation to arbitrariness. Later, in *Gregg v. Georgia*, a Georgia statute which permitted death sentences to be imposed under a statutory system characterized by "guided discretion" was endorsed, in principle, as a means to rein in otherwise untrammeled discretion. This precedent, which affected statutes in a number of states, endorsed statutes that delimited jury sentencing discretion by requiring, at a minimum, that the jury find at least one "aggravating factor" from a legislatively endorsed list of aggravating factors before they were authorized to sentence a convicted offender to death. Once found, however, other kinds of aggravating factors could be considered, not just those listed by statute.

It is, of course, one thing to proscribe arbitrariness in principle and quite another to prevent it in practice. Significantly, there have been a very large number of recent studies investigating the role of inappropriate factors in capital sentencing (Selin 1980; Baldus et al. 1985; Paternoster 1983; Radelet & Pierce 1985; Barnett 1985; Paternoster & Kazyaka 1988; Gross & Mauro 1989; Keil & Vito 1989). A review of these and other studies by the U.S. General Accounting Office (1990) led to the conclusion that efforts by the states to provide effective procedural safeguards had failed. In jurisdiction after jurisdiction, the victim's race was a statistically important factor in determining which offenders were sentenced to death, controlling for certain other variables. In particular, the victimization of whites was treated far more seriously than that of blacks, so that individuals convicted of murdering whites were at greater risk of a death sentence than individuals convicted of murdering blacks, even if the seriousness of the crimes and prior criminal records were comparable.

These conclusions, whatever their merits, may well be legally irrelevant. Five years ago, the Supreme Court of the United States ruled (*McCleskey v. Kemp* 1987) that aggregate findings of discrimination have no constitutional bearing on whether a given individual has been discriminated against. To

assert that "on the average," for example, offenders who murder whites are four times more likely to be sentenced to death than comparable individuals who murder blacks does not conclusively resolve whether the offender in question was discriminated against (even if his/her victim were white).

In this article, therefore, we turn to the second kind of arbitrariness: *capriciousness*. In the *Furman* decision, Justice Stewart observed that existing jury sentencing was "cruel and unusual in the same way that being struck by lightning is cruel and unusual" (*Furman v. Georgia* 1972:209). Justice Brennan agreed that the existing procedures were "little more than a lottery system" (p. 293). Justice White found "no meaningful basis for distinguishing the few cases in which [the death penalty] is imposed from the many in which it is not" (quoted in Gross & Mauro 1989:6). Such problems were supposed to have been eliminated by current death penalty statutes, found constitutional under *Gregg*, which ostensibly constrained and guided the sentencing discretion of prosecutors and capital juries.

It is, therefore, constitutionally important to know whether, in fact, the revised state death penalty statutes are effectively reducing or eliminating capriciousness in capital sentencing. For as Justice White observed in *Furman*, " 'legislative policy' is . . . necessarily defined not by what is legislatively authorized, but by what juries and judges do in exercising the discretion so regularly conferred upon them" (*Furman* 1972:314). To answer part of this broader question, we address below one important step in the death sentencing process: the prosecutor's decision to charge a offender with a death-eligible homicide. We begin by exploring how the concept of a lottery may be used to inform the justices' concerns about capriciousness. Then, using data from 1978 through 1988 for San Francisco, we illustrate how one might empirically address the question of capriciousness; after briefly examining arbitrariness through discrimination, we consider arbitrariness through capriciousness. Finally, we speculate a bit about the implications of our illustration for the conception of capriciousness.

It is perhaps important to stress that we will take an unorthodox stance. In particular, we are ultimately concerned with *how the probabilities of a capital charge are distributed across offenders in a given jurisdiction*. For example, one state's criminal justice system might effectively group homicide offenders into two categories. Most fall into a category in which the probability of a capital charge is very small, perhaps 1 in 500. But a few fall into a category in which the probability of a capital charge is rather high, perhaps 1 in 2. In contrast, another state's criminal justice system might effectively group homicide offenders into a large number of categories in which the likelihood of a capital

charge clusters tightly around a central tendency of about one capital charge for every 20 homicide cases.

We argue below that it is in the consideration of outcome distributions such as these that *capriciousness* can be usefully examined. While we make a serious effort to uncover the determinants of capital charges, we do not here build our substantive story on the specific characteristics of offenders, victims, and cases that lead to particular charging distributions; we are far less interested in why certain explanatory variables may prove to be important than in capriciousness per se.

In summary, our article has two primary goals. The first is to raise for consideration how one might constructively think about the role of chance in the processing of homicide cases. The random component is *not* a statistical approximation of what is really unfolding but is for practical purposes the empirical reality itself. The second is to highlight the distribution of charging outcomes as a critical jurisprudential issue in its own right. Far more is involved than statistical arguments about the amount of "explained variation" or the "goodness of fit."

I. Some Conceptions of Capriciousness

Perhaps the only common theme in the Justices' references to capriciousness has been overwhelming uncertainty. In *Furman*, it was, in their view, impossible meaningfully to distinguish individuals who were sentenced to death from those who were not. One way to add more conceptual precision to this observation is to take seriously Justice Brennan's allusion to a lottery.

None of the justices believed that death penalty decisions were literally lotteries. That is, there was no conscious effort by criminal justice decisionmakers to produce a stochastic outcome, as one might by flipping a coin, spinning a roulette wheel, or drawing a card from a thoroughly shuffled deck. Rather, the Justices were concerned that the *Furman*-era statutes inadvertently produced outcomes that looked *as if* they had been—and may equally well have been, in fact—produced by lottery mechanism. Justice White expressed the Supreme Court's hope that the appearance of a lottery-like process would be overcome by the new "guided discretion" capital statutes like those conditionally approved in *Gregg* (1976:222):

> As the types of murderers from which the death penalty may be imposed become more narrowly defined and are limited . . . by reason of the aggravating circumstances requirement, it becomes reasonable to expect that juries—even given discretion *not* to impose the death penalty—will impose the death penalty in a substantial portion of the cases so defined.

If they do, it can no longer be said that the penalty is being imposed wantonly or freakishly.

A One-Urn Model of Charging

What would an "as if" lottery look like? Consider first a situation in which there is a pool of offenders arrested for homicide. Consider also a single urn with three red balls and seven white balls. For each offender in the pool, suppose the balls are thoroughly stirred and a single ball blindly picked. If the ball is red, the offender is charged by the prosecutor with a death-eligible homicide; that is, if convicted, the offender could face the death penalty. If the ball is white, the offender is not charged with a death-eligible homicide but with some lesser degree of homicide. The key idea is that each offender is subjected to exactly the same probability of a death-eligible charge, and in that sense, each offender can be considered "exchangeable."

Suppose, now, that in a particular state jurisdiction the proportion of people charged with a death-eligible homicide is 30%. Moreover, suppose that after a careful examination of the facts surrounding each case, no subsets of offenders can be found whose conditional probability of a death-eligible charge is systematically different from 30%. It might then be reasonable to conclude that the real-world charging system behaved as if well-mixed red and white balls were being pulled from a single urn. At least, this seems consistent with the Justices' formulations.

Three points need to be stressed. First, as the *Furman* decision illustrates, *uncertainty is in the eye of the beholder*. It is always theoretically possible that apparently random differences in the judicial treatment of death-eligible cases reflect some systematic treatment of those cases by some as yet unidentified variable. Therefore, the assurance that we are dealing with "as if" capricious sentencing outcomes depends upon the care with which one has previously identified systematic patterns in the data. Observers may differ, therefore, on whether sufficient detective work has been undertaken and on whether a convincing case has been made.[1] We will later confront these issues directly when our empirical example is discussed.

Second, self-interested offenders would be *indifferent* between such an "as if" lottery and the real-world charging system. The risks of an adverse *outcome* would be identical under

[1] To some, this will seem like scientific heresy. For them, there is a single objective reality, and with proper tools and sufficient diligence, that single reality will be revealed. Here, we take no explicit position on such issues, which are well beyond the scope of the article. Our point is that as a *practical matter* one does the best one can with the resources available in this imperfect world in which assessments must be made.

both. Therefore, from the point of view of the self-interested offender, the two charging schemes would for all practical purposes be identical. Put another way, since the primary concern of a death-eligible capital defendant is whether or not a capital sentence is actually imposed, from the self-interested offender's point of view—viewed strictly from an outcome perspective—the real-world charging system is no different from a lottery.

Finally, should a particular jurisdiction produce a set of charging decisions consistent with a one-urn "*as if*" lottery, important constitutional issues would be raised. In effect, those charging decisions would be fully determined by nothing different from what amounts to the luck of the draw. Moreover, a constitutional challenge to such sentences should not be vulnerable to the objection interposed by the Supreme Court against the statistical evidence of racial discrimination presented in *McCleskey*.

In that case, Justice Powell held that "to prevail under the Equal Protection Clause (of the Fourteenth Amendment), McCleskey must prove that the decision-makers in *his* case acted with discriminatory purpose" (*McCleskey v. Kemp* 1987:292; emphasis in original). In other words, absent affirmative evidence by McCleskey that his own prosecutor or his jury engaged in intentional discrimination in disposing *his own case*, no general pattern of discriminatory capital sentencing, however strong, would suffice to warrant a finding that his death sentence was constitutionally infirm. Later in his opinion for the court, Justice Powell rejected McCleskey's alternative constitutional challenge—asserted under the cruel and unusual punishment clause of the Eighth Amendment—because McCleskey had shown no more than that particular Georgia inmates convicted of murder under similar circumstances had not received death sentences. "*[A]bsent a showing that the Georgia capital punishment system operates in an arbitrary and capricious manner*, McCleskey cannot prove a constitutional violation by demonstrating that other defendants who may be similarly situated did *not* receive the death penalty" (*McCleskey* 1987:306-7; emphasis added).[2]

Recall also the words of Justice White (quoted earlier) who found "no meaningful basis for distinguishing the *few* cases in

[2] Justice White similarly indicated in *Gregg* that an empirical showing might be proffered to substantiate a constitutional attack based on arbitrary or capricious sentencing patterns:

> Petitioner's argument that prosecutors behave in a standardless fashion in deciding which cases to try as capital felonies is unsupported by any facts. Petitioner (Gregg) simply asserts that since prosecutors have the power not to charge capital felonies they will exercise that power in a standardless fashion. This is untenable. *Absent facts to the contrary*, it cannot be assumed that prosecutors will be motivated in their charging decisions by factors other than the strength of their case and the likelihood that a jury would impose the death penalty if it convicts. (*Gregg* 1976:255; emphasis added)

which [the death penalty] is imposed from the *many* in which it is not" (emphasis added). In other words, because the Court must necessarily rely on comparisons between sets of cases in order to assess the presence of possible capriciousness, statistical evidence is clearly relevant; McCleskey-like reasoning for disregarding statistical evidence would seem not to apply. Put more strongly, irrespective of the Court's views, *any* meaningful empirical determination of whether or not an "as if" lottery exists would seem to require data on a number of offenders.

A Many-Urn Model of Charging

Consider a slightly more complicated process than our one-urn model. There are now several urns with red and white balls. The urns differ in the fraction of red balls and in the subpopulation of offenders whose fate they determine. For example, offenders who have committed especially heinous homicides and have especially serious prior records are assigned to an urn with nine red balls and one white ball. At the other extreme, offenders whose homicides are not highly aggravated and who are first offenders are assigned to an urn with one red ball and nine white balls. Offenders between the extremes are assigned through legitimate factors to urns with more balanced mixtures of red and white balls. Legitimate factors might include the existence of provocation, whether there was premeditation, and the weapons used. In short, each of several kinds of offenders is assigned to an urn with a given conditional probability of drawing a red ball and, therefore, a death-eligible charge.

Returning once again to the "real world," imagine that a data analysis of actual charging decisions produced estimates of conditional probabilities that were identical to those given by the urns. That is, each offender would seemingly face the same conditional probability under both systems: the hypothetical lottery and the charging system in practice. Moreover, imagine that careful scrutiny of the data unearthed no variables of any kind to distinguish those individuals within each urn charged with a death-eligible offense from those not charged with a death-eligible offense.[3] Each collection of similarly situated offenders was subjected to the same (and only one) conditional probability; for each collection of offenders, a particular Bernoulli process apparently held (and not a mixture of Bernoulli processes). Therefore, within each collection of similarly situated offenders one has, once again, an "as if" lottery.

[3] As noted already above, we believe that this is the best one can do. There currently exists no *definitive* means to determine whether one has the story right. More fundamentally, the idea that there is a single "right" story is perhaps problematic to begin with.

It is impossible to know what the Supreme Court Justices would do with this analysis. The many-urn lottery is some distance from the clearly objectional single-urn lottery and yet admits a significant role for chance. Much depends on the implications of different conditional probabilities: how many urns and the *shape* of the distribution of conditional probabilities.

Using a simple example for illustrative purposes, a particular "as if" system might be acceptable if it were characterized by a large number of urns with most of the conditional probabilities concentrated near zero or one. Thus, if 90% of all death-eligible defendants included in the urn for highly aggravated homicides were charged with a capital offense, this might be constitutionally acceptable. Extrapolating somewhat from statements by Justices Stewart, Powell, and Stevens in *Gregg* (1976:203), "the isolated decision of a jury to afford mercy does not render unconstitutional death sentences imposed on defendants who were sentenced under a system that does not create a substantial risk of arbitrariness or caprice." Implied is that there is really a set of "good guys" and a set of "bad guys" who are appropriately subjected to very different risks of death-eligible charging, and that, stated another way, most of the variability in charging can and should be attributed to which urn offenders were assigned.

In contrast, another "as if" system might be unacceptable if it were characterized by a very small number of urns with conditional probabilities all clustered around .50. Thus, if 50% of all convenience store robbers with similar prior records and characteristics were changed with a death-eligible offense for shooting a storekeeper in such robberies, this would not be consistent with an "isolated decision . . . to afford mercy" to one of every ten offenders, but instead is a virtual 50–50 coin flip. Implied is that meaningful distinctions between offenders were difficult to make and that most of the variability in charging could be attributed to the Bernoulli processes.

More challenging, and far more interesting, would be distributions that better mirrored real-world jurisdictions with, for example, several modes or large gaps, or long tails. In other words, beyond summary statistics such as the mean and variance, the Justices would have to consider the *form* of the conditional probability distribution. For example, suppose in one jurisdiction fine distinctions were made between offenders so that the proportion of offenders charged with a capital crime effectively varied between 0.0 and 0.70, but with each of the many categories having about the same proportion of offenders (i.e., a rectangular distribution between 0.0 and 0.70). Suppose in another jurisdiction the proportion of offenders charged with a capital crime effectively varied between 0.0 and 1.0, with about a two-thirds of the offenders clustered near 0.0 and with

the remaining third spread out evenly over the many categories represented by rest of the range (i.e., a long tail to the right). The two distributions could have approximately the same mean and variance but imply very different patterns of charging outcomes. Which would be more capricious?

Individualized Deterministic Charging

Imagine that each charging decision is made deterministically based on a host of case-specific variables: the weapon used, provocation by the victim, the offender's prior crimes, the kinds and strength of forensic evidence, degree of premeditation, and so on. The decision would be deterministic in the sense that, were it possible for different prosecutors (or the same prosecutor) to review the same case over and over independently, the same outcome would result each time.

This is, of course, the kind of individualized justice in charging (and sentencing) that the Supreme Court has applauded. The prosecutor takes into account a very large number of legitimate variables, many with values that are unique for the case at hand, and then makes an unambiguous decision; each case is evaluated thoroughly on its particular merits. There are no rules of thumb and no shortcut decision procedures.

However, the Court's ideal leads to serious conceptual and practical problems. Even if deterministic individual charging were universal, a *prospective* examination of the charging process would necessarily produce an "as if" lottery experience for the defendant, the defense attorney, and even the prosecutor. If each case is truly unique, *past decisions cannot be used to predict with certainty how the new charging decision will come out*. It is impossible to know in advance exactly which characteristics of the crime and the defendant will (and will not) be used in the charging decision and how the selected characteristics will be aggregated to arrive at an overall judgment. In short, it is impossible to predict with certainty the future from the past, since the defendant's particular "decision rule" has yet to be written.

A different set of complications surface in any *retrospective* attempt to account for a given charging decision. Presumably, a decision rule has been applied, but the task now is to reconstruct it so that a fully accurate "backcast" can be made. That is, an observer not knowing the actual charge could reproduce the charge retrospectively with perfect accuracy. However, if each case is really unique, the exact means by which the charging decision was made must be unique as well; for every thousand defendants, there are a thousand unique decision processes, all of which would have to be recorded so that a later reconstruction is possible. That reconstruction, in turn, would

have to capture with complete accuracy the factors that were taken into account and how those factors were combined into an overall charging decision. The only observer who could conceivably have full access to this information would be the prosecutor directly responsible for each decision; much of the relevant information would have to be recalled from memory. All other observers—the defendant, the defense attorney, the sitting judge(s)—could not possibly know all the relevant factors and how they were used. The resulting uncertainty would necessarily create an "as if" lottery for these observers.

What about the prosecutor directly responsible for the cases in question? Perhaps the most important point is that there is no means by which any prosecutor's retrospective account of the charging decision could be fully evaluated; no one else could possibly have all the necessary information. For example, suppose a defendant is charged with a death-eligible homicide despite the absence of a serious prior record or any compelling aggravators. To some outsiders, a "lightning strike" may have occurred. But, suppose the prosecutor responds by claiming that any possibility of leniency was ruled out because the homicide was committed in a particularly callous manner. There is no external information by which the prosecutors claim may be corroborated. One might be able to learn something of the defendant's motives, but how could one reconstruct from external sources what use the prosecutor made of that information? In short, if each case is truly unique, important parts of prosecutors' retrospective accounts cannot possibly be impeached.

Of course, real-world prosecutorial charging is very little like the individualized deterministic ideal. To begin, years of research in psychology on expert and clinical judgments make clear that for any given case, a prosecutor could not possibly take into account the amount of information the ideal requires. *Indeed, the upper boundary for the number of factors that could be simultaneously and reliability evaluated is probably well under 10* (Faust & Ziskin 1988). Moreover, those factors that are taken into account are potentially subject to a wide variety of nontrivial distortions (Tversky & Kahneman 1974). In other words, the individualized deterministic ideal places unrealistic demands on human cognitive abilities, and "as if" lotteries would seem to necessarily follow from inherent limitations in the amount of information that the human mind can reliably manipulate.

It is also widely recognized that the organizational settings in which prosecutors are likely to work lead to rules of thumb, shortcut decision procedures, and numerous uncertainties (Neubauer 1974; Rosett & Cressey 1976; Stanko 1981; Maynard 1984; Sanders 1987). Thus, for many cases, there are relatively few hard and fast rules linking particular offenses and of-

fenders to particular charges; much depends on myriad "judgment calls." Where exactly, for example, is the line that separates the very heinous homicide for which a death-eligible charge may be appropriate from the only somewhat heinous homicide for which a death-eligible charge may not be appropriate? Would a different prosecutor looking at the same case make the same decision; indeed, if the same prosecutor could evaluate the same case again with the prior judgment erased from memory, would the same conclusion result?

This is not to disparage the efforts of prosecutors. We are simply acknowledging the obvious: fitting the offender to the charge often takes judgment and wisdom, producing at best loose approximations of the charging ideal, which cannot be fully replicable. This does not mean, however, that all cases would fail the test of replicability. Many would pass, because at the extremes, charging decisions are often "overdetermined." The nature of the crime and the background of the offender both point clearly in the same direction. For example, individuals who have already committed murders or rapes and who then commit double or triple torture-style homicides are almost always charged with capital crimes.

To summarize, the many-urn model may be a reasonable description of charging decisions in homicide cases not just because of ignorance on the part of outside observers, and even prosecutors themselves, but also because all charging decisions cannot be deterministic in practice. We are suggesting, therefore, that prosecutorial decisions to charge suspects with a capital crime or to prefer "special circumstances" may be *empirically indistinguishable* from "as if" lotteries, as we conceptualized them. This implies that it should be possible to formulate statistical models of our "as if" lotteries using data on actual charging decisions that could be used to address capriciousness in particular jurisdictions. Moreover, such work could be used to underscore the question of which conditional probability distributions for charging may be acceptable to the courts. Indeed, it is the careful examination of the conditional probability distributions that we are especially advocating.

Four points should perhaps be stressed. First, even if a particular statistical model fits the data well, it does *not* mean that the model is correct in a causal sense; the actual *mechanisms* by which charging decisions are made may not be properly represented. It only means that offenders can be sorted into groups, each group with a single conditional probability of a capital charge. The residual heterogeneity, no doubt explained by a large number of omitted variables, produces in the aggregate a distinct binomial "experiment" for each cluster of offenders, each with a single probability of a death-eligible charge.

Second, these statistical requirements for a good fit are very

demanding. While any empirical efforts are necessarily constrained by the amount of information in a given data set, evidence must be provided that a good fit has in fact been achieved. A simple assertion that all is well will not suffice.

Third, other models, perhaps implying very different causal mechanisms, may also fit the data. Yet, a good fit between an "as if" lottery model and real charging decisions is all we need to make our case that an "as if" lottery is operating in a particular jurisdiction; we require only that systematic patterns in the data have been accounted for, that no models fit better, and that there is residual variation that will not go away.

Fourth, with the existence of an "as if" lottery empirically established, concrete considerations of the *nature* of the capriciousness require that the model's conditional expectations be carefully examined. This is where the important story lies. The enterprise we are proposing is a bit like forecasting. Good forecasts are often defined in terms of their predictive distribution and their relationships to the "true" (but initially unobserved) values to be predicted. Minimizing the sum of the squared forecasting errors is one illustration. There is nothing in this definition that speaks to the information used to make the forecast. The focus is on "output" not "input."

II. An Illustration

The Data

To illustrate how various conceptions of arbitrariness might apply to a real-world setting, we turn to data on all homicides in the County of San Francisco from 1978 through 1988. The data were coded by staff in the Public Defender's Office from official records and forms filled out by police and prosecutors. Vehicular homicides were later excluded because, a priori, they are not death eligible under California law. In order to keep the analysis logically neat and avoid the selection biases that can result from sampling on the outcome variable, such cases were excluded solely on the basis of information that preceded the charging decision.

The outcome of interest is the decision by prosecutors to charge offenders with "special circumstances." In California, this means that the case is deemed to have certain aggravating characteristics specified by statute, which if affirmed either by plea or by trial would make the offender eligible for a death sentence. In short, if an offender is charged with special circumstances, the prosecutor is seeking the death penalty.

Of the 363 homicide cases in our data set, 7.4% (27 defendants) were charged with special circumstances. Clearly, prosecutors in San Francisco (and elsewhere) seek the death penalty

Table 1. Logistic Regression Results

Variable	Parameter Estimate	Standard Error	LRT Statistic	*P*-Value	Range of Variable	Odds Multiplier
Constant	−10.70	1.536	—	—	—	—
Offender White	1.569	0.6236	6.76	.0093	0–1	4.80
Victim Female	1.298	0.6167	4.19	.0407	0–1	3.66
No. prior serious felonies	0.6817	0.1696	16.8	.0000	0–8	1.98
No. prior homicides	2.928	0.7247	17.1	.0000	0–2	18.7
No. victims	4.593	0.8382	46.7	.0000	1–4	98.8
Victim stranger	2.119	0.6346	12.7	.0003	0–1	8.32

NOTE: *P*-values are from the likelihood ratio test (LRT) for testing the parameter equal to zero. These are asymptotically distributed as a χ^2 with 1 degree of freedom under the null hypothesis that the parameter is equal to zero. We do not test the constant term for significance.

in only a small fraction of homicide cases. This is both a finding of some interest and a potential problem for data analysts. With such a small fraction of offenders being charged with special circumstances, it is difficult to consider simultaneously the impact of more than a few explanatory variables.

Over 70 explanatory variables were available addressing such considerations as the number of victims, the relationship between the victim and the offender, the means of killing, possible motive, other crimes that may have been committed contemporaneously, the background of the offender and victim and so on. At the same time, however, past research (cited above) has shown that much of the story can be told with a modest number of explanatory variables. For example, Gross and Mauro (1989) build their fine book around analyses in which generally fewer than a half-dozen explanatory variables are used. We argue below that for our data, including more than a very modest number of variables does not demonstrably improve the fit.

The Statistical Analysis

One convenient way to translate our many-urn model into a means of analysis is to apply logistic regression, traditionally used in the analysis of death penalty sentencing. The explanatory variables in the logistic regression reveal, in effect, which urn was used for which offender, while the predicted probability for each offender would represent the fraction of balls in the urn that were red. Table 1 shows the explanatory variables on which we will concentrate. The first two variables represent illegitimate factors (race and gender), and the rest represent factors that speak to the seriousness of the crime or the prior record of the offender. With the exception of "Offender White," all of the variables were chosen *before* any mod-

els were examined by drawing on the death penalty empirical literature, which suggests the factors most associated with sentencing outcomes in homicide cases.[4]

Because several of the explanatory variables were quite skewed, we also estimated later some models with the explanatory variables recoded to reduce the leverage of certain observations (e.g., recoding all of the explanatory variables to dichotomous 0–1 variables). In addition, efforts were made to add variables to the model that past research had not considered (e.g., whether the victim was gay) or variables that were effectively alternative indicators of the same phenomenon (e.g., the total number of prior felony convictions rather than only the number of serious prior felony convictions). Perhaps most important, we were trying to find alternative, legally acceptable, and theoretically plausible explanations for apparent racial and gender effects. We found none, and the overall story did not change. In addition, we were, in an exploratory manner, trying to enrich the level of explanation. However, here too nothing important was altered. We will be more specific shortly.

The substantive story in Table 1 is extracted from the odds multipliers (which are statistically significant by likelihood ratio tests as well). Clearly, some of the effects are very large. For example, the odds of a death-eligible charge for white offenders are about five times larger than for nonwhite offenders, other things being equal. The odds for individuals who kill two people (about 3% of the sample) are nearly a hundred times larger than for people who kill only one person.

One way to explore the credibility of our results is to apply the diagnostic procedures developed by Landwehr et al. (1984). In effect, one *simulates* what the residuals from the logistic regression would look like if the model were correct (a Bernoulli process with a single predicted probability for each configuration of explanatory variables) and then compares the simulated residuals with the actual residuals (see Appendix). Possible problems with the model are indicated by large disparities between the actual and simulated residuals (see below). When no such disparities are found, it increases one's confi-

[4] "Offender White" refers only to the suspect's race and, given the literature, perhaps requires a bit of discussion. Ideally, we would have preferred to consider simultaneously the impact of race of suspect, race of victim, and their interactions. However, there are four large racial groups in San Francisco (whites, blacks, Latinos/as, and Asians), well beyond what could be explored with our highly skewed outcome variable. At first, we examined charging effects attributable to the victim's race, in part because these are commonly found in the literature. Although there was some evidence of race-of-victim effects, particular racial effects were very difficult to isolate. We finally settled on the main effect of "Offender White," because it had perhaps the greatest impact among the racial variables. Recall, however, that if one is concerned with capriciousness, such matters are secondary.

dence in the model, albeit without supplying definitive proof that the model is "correct" (Flack & Flores 1989).

Consider Figures 1 and 2. Figure 1 shows the diagnostic plot for the model reported in Table 1. If the simulated and actual residuals are much the same, a plot of one against the other should produce nearly a straight line through the origin, 45 degrees from the horizontal axis. In fact, most of the points fall very near the straight line, and with about a half-dozen exceptions, all of the points fall on or within what Flack and Flores (ibid., p. 220) call the "diagnostic envelope."

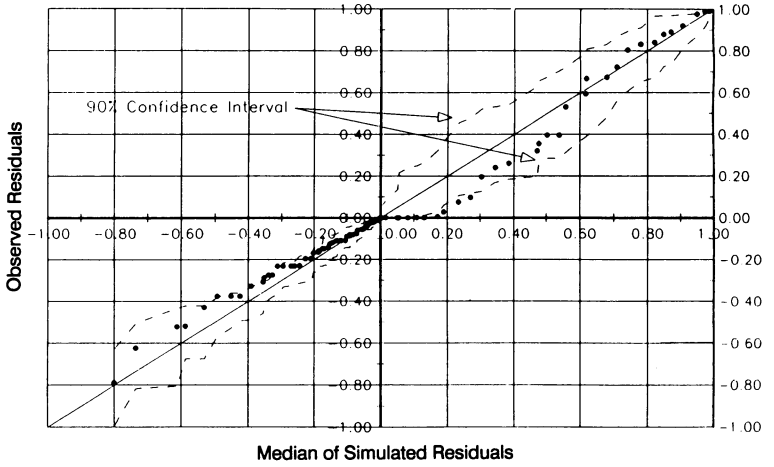

Figure 1. Diagnostics for baseline model

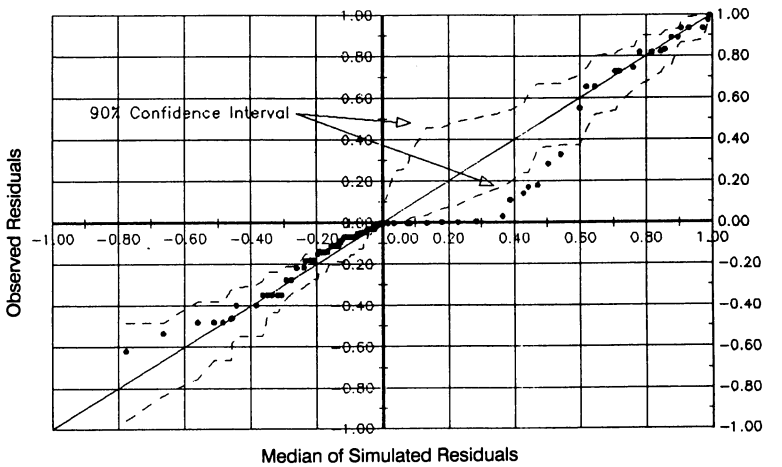

Figure 2. Diagnostics for model with race and gender removed

https://doi.org/10.2307/3053749 Published online by Cambridge University Press

The small cluster of potentially problematic points near the center of the graph reflect a few cases in which offenders were charged with special circumstances and where our predicted probability of that charge was very low. But by and large, the model does rather well. When "white offender" is replaced by an interaction variable, for white-suspect-and-white-victim, the results reported in Table 1 for the other explanatory variables were substantively identical and the fit to the simulated residuals virtually the same as in Figure 1. That is, at least two models are consistent with our "as if" many-urn formulation. In contrast, Figure 2 shows the same simulation with the race and gender variables removed from the model. Clearly, far more points fall some distance from the straight line and outside the diagnostic envelope. Thus, one is able to discard some models.

While there is certainly no means to prove that our model is "right," its overall credibility was systematically explored in two other ways as well. First, a quantile-quantile display was constructed in which a set of χ^2 statistics for adding each potential additional variable was plotted against the quantiles from the χ^2 distribution. A close approximation of a straight line resulted. That is, the larger χ^2 values observed were well within what one would expect in a number of independent draws from a single χ^2 distribution.

Second, we constructed added variable plots. Figure 3 shows the effect on the fitted probabilities of adding in the variables Offender White and Victim Female (when no other race or gender variables are already included—i.e., the model in Table 1 without race and gender). The horizontal (x) axis is the estimated probability under a model that omits Offender White and Victim Female. The vertical (y) axis gives the change in the estimated probability after adding in these variables. Circles indicate offenders charged with special circumstances, points indicate offenders not charged. The change in probability is largely positive for the circles. That is, approximately 12 circles are substantially above the line at $y=0$, indicating better fits for 12 cases, 4 circles are substantially below the line, indicating worse fits for 4 of the death-eligible offenders, and not much change for 11 cases quite near the line. It appears, therefore, that race and gender, by and large, improve the goodness of fit in sensible ways. In contrast with perhaps one exception, *none* of the other variables in our data set seemed to be useful beyond what would be expected by chance alone. And that one variable made no substantial difference in the fit.[5]

[5] That variable was whether the police report that the victim was "bound and gagged." However, a number of other variables also related to whether the crime was especially heinous, and none of these made an incremental difference in the model. Part of the problem was that the number of cases falling in these categories was very small. Another problem was the accuracy of the police accounts for events that might be difficult to verify, such as whether the victim pleaded for mercy.

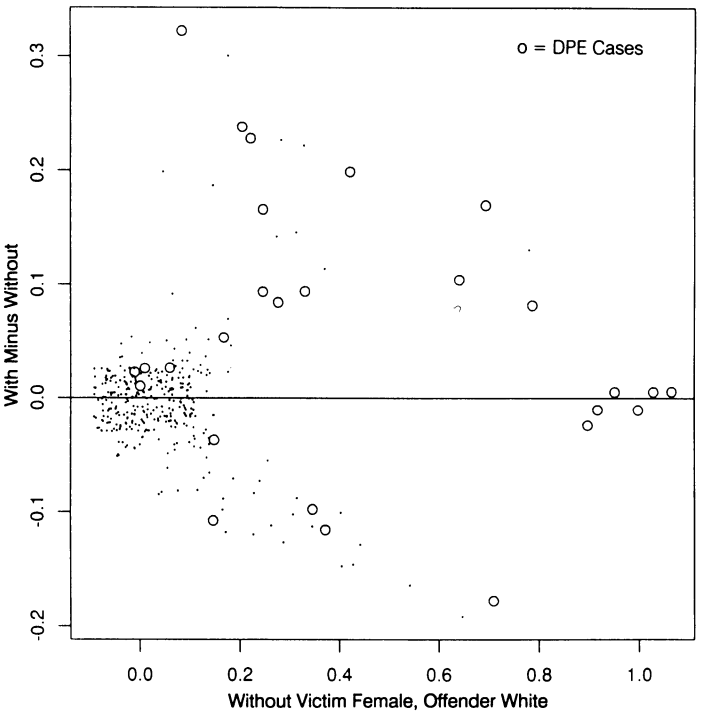

Figure 3. Change in fitted probabilities due to victim female and defendant white variables.

There are at least two ways in which these analyses may be construed. First, the model diagnostics may be taken at face value as goodness-of-fit indicators. As such, they provide some comfort but certainly do not require that our models are properly specified in a causal modeling sense. Second and far more important, the Landwehr et al. diagnostic is based on a *simulation* of the many-urn lottery described earlier. In other words, the computer algorithm described in the Appendix, produces the equivalent of a distinct Bernoulli process for each cluster of offenders identified by the logistic regression. This means that a computer simulation of a particular many-urn lottery is consistent with our statistical model of how a set of charging decisions was actually made. Put more strongly, *the lottery mechanism represented by our logistic regression analysis of data from San Francisco is empirically indistinguishable from our computer-generated lottery mechanism.* Arguably, therefore, we meet our "as if" requirement.

Finally, if our results are consistent with a particular many-urn lottery, what does the output of that lottery look like? Recall our earlier argument that it was through the distribution of outcomes that capriciousness might usefully be considered.

https://doi.org/10.2307/3053749 Published online by Cambridge University Press

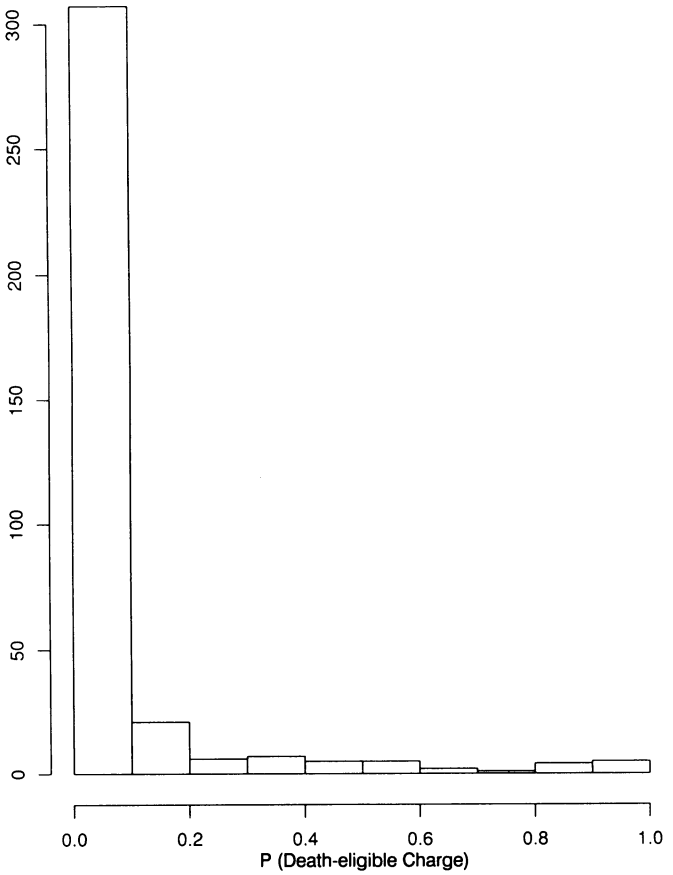

Figure 4. Histogram of fitted probabilities

Figure 4 shows a histogram of the fitted conditional probabilities of a capital charge, given our data and model.

Clearly, for the vast majority of homicide offenders, the risks are very low, and few important distinctions are made between these offenders. For all practical purposes, these offenders are in all the same urn. But because the probability of a capital charge is nonzero, one offender is occasionally charged with a capital crime. From the point of view of that unlucky offender, the charging system must seem very capricious indeed! But since our concern is with the charging system, we must focus on the distribution as a whole. Figure 4 shows that there is also a small group of offenders whose risks are spread across the moderate range and a very small group of offenders for whom the risks are extremely high. In other words, *the distribution of conditional probabilities is highly skewed to the right.*

What does Figure 4 convey about systemic capriciousness? We have argued that capriciousness in the criminal justice system is closely related to an inability to predict, either prospec-

tively or retrospectively, a criminal justice outcome. An inability to predict, in turn, is nothing more than a difference between an outcome projected from existing information and the actual outcome. A conception of systemic capriciousness might then combine one or more measures of the gap between the projected and actual outcome (i.e., a deviation score) with one or more ways of aggregating those over offenders. Stated more formally, capriciousness would need to be addressed through loss functions.

For example, drawing from common measures of spread, it would be natural to consider some function of the squared deviations summed over offenders. However, we know of no jurisprudential justification for this approach. In addition to weighting an overall measure of capriciousness heavily in favor of the larger disparities, the sum of squared deviations assumes symmetry in the weighting. In particular, a "lightning strike" capital charge among offenders near the left tail of the distribution counts the same as a "lightning strike" reprieve from a capital charge among offenders near the right tail of the distribution. This would equate unforeseeable leniency with unforeseeable punitiveness. Our general point is that the selection of any summary measure of capriciousness requires further thought and that it is best, therefore, to begin with a consideration of the full distribution. What one then makes of that full distribution depends on how one weighs its different parts.

III. Discussion and Conclusions

If one were worried about omitted variables for the usual reasons, our formulations would nonetheless have more credibility than most that have examined death sentencing in the past. In addition to essentially replicating much past research, the diagnostics, model explorations and simulations we have employed provide some useful supporting evidence, heretofore unavailable, for the model. However, ours is not the only model that might survive the diagnostic analysis, and although race and gender appear to be important, there may be omitted variables with which they are highly correlated (and which have the same pattern of correlations with the other explanatory variables) that could reproduce the apparent effects. And these variables could conceivably be legally permissible.

Yet, here we are far more concerned with capriciousness than with the content of particular explanatory variables. The logistic regression diagnostics developed by Landwehr et al. (1974) allowed us to see how consistent our model of a particular set of charging decisions is with a many-urn lottery. It appears to us that the fit is quite good. This would seem to imply that death penalty charging decisions in San Francisco meet

our "as if" lottery criterion. It does not matter that one other (at least) model might fit as well. What we are saying is that for many observers, including us, the charging process looks stochastic. We are also arguing that a self-interested offender would be indifferent between the "as if" lottery and the actual charging process because the charging outcome would be virtually identical. In that sense, both are equally real.

More important, we have shown that the "as if" lottery of charging decisions produces a highly skewed distribution of charging risks. Clearly, a large number of other distributions might have been produced: a bimodal distribution with cases clustered at the extremes, a unimodal distribution skewed to the right, a bell-shaped distribution centered at moderate levels of risk, a relatively flat distribution across the entire range of risk, and so on. Each of these distributions describe different patterns of the uncertainty for death-eligible charging. And if uncertainty is at the heart of jurisprudential concerns about capriciousness, these patterns may well contain very critical information; they provide information not just about the amount of uncertainty overall *but how the risks are spread across different kinds of offenders*.

For example, suppose that in another jurisdiction the "as if" charging lottery produces a bell-shaped distribution of conditional probabilities centered on a conditional probability of about .25. Suppose also that the range of this distribution fell between .00 and .50 and that its variance was approximately the same as the highly skewed distribution shown in Figure 4. Which "as if" lottery is more capricious? Extrapolating from Justice White's apparent interest in being able to distinguish between the few at risk to a death sentence from the many not at risk, the highly skewed distribution is perhaps less capricious. However, any overall assessments of systemic capriciousness require a loss function. One could, in principle, construct a loss function that would make the highly skewed distribution more capricious.

The point is that any (and all) specifications consistent with a many-urn lottery would mean that current Supreme Court conceptions of lotteries, lightning strikes, or inexplicable sentencing patterns are at best incomplete. It seems to us that "as if" lotteries are virtually inevitable in death penalty charging (and all criminal justice processes more generally). "As if" lotteries are not a mere nuisance, which with proper procedure and vigilance can be effectively eliminated. Rather, they are an inherent and inevitable part of the whole. The critical job, therefore, is to figure out which sorts of "as if" lotteries are constitutionally acceptable and which sorts are not.

Appendix: Simulation Steps

1. From the fitted model: logit $(\hat{p})=X\hat{\beta}$ obtain residuals $r_i=y_i-\hat{p}_i$, $i=1, \ldots, N$.
2. Order the r_i, giving $r_{(i)}$.
3. Simulate data y^* from the fitted model, i.e., $y_i^* \sim$ binomial $(1, p_i)$, $i=1, \ldots, N$.
4. Fit the model: $(p^*)=X\beta^*$ to these data.
5. Compute the residuals $r_i^*=y_i^*-p_i^*$ and order them, giving $r_{(i)}^*$
6. Repeat steps 3–5 M times independently.
7. Compute the typical values (e.g., medians) of the ordered residuals over the M replications, say, $T_{(i)}=\text{med}\ \{r_{(i)}^*\}$, $i=1, \ldots, N$.
8. Plot the ordered residuals from the original fit against the typical ordered residuals from the simulation, i.e., plot $(T_i, r_{(i)})$, $i=1, \ldots, N$.
9. Obtain and plot upper and lower confidence bounds at each of the $T_{(i)}$ by taking upper and lower quantiles of the M values $\{r_{(i)}^*\}$ corresponding to the desired confidence coefficient.

References

Baldus, David C., George Woodworth, & Charles A. Pulaski (1985) "Monitoring and Evaluating Contemporary Death Sentencing Systems: Lessons from Georgia," 18 *U.C. Davis Law Rev.* 1375.

Barnett, Arnold (1985) "Some Distribution Patterns for the Georgia Death Sentence," 18 *U.C. Davis Law Rev.* 1327.

Faust, David, & Jay Ziskin (1988) "The Expert Witness in Psychology and Psychiatry," 241 *Science* 31.

Flack, Virginia F., & Rafael A. Flores (1989) "Using Simulated Envelopes in the Evaluation of Normal Probability Plots of Regression Residuals," 31 *Technometrics* 219.

Gross, Samuel R., & Robert Mauro (1989) *Death & Discrimination: Racial Disparities in Capital Sentencing*. Boston: Northeastern Univ. Press.

Keil, Thomas J., & Gennaro F. Vito (1989) "Race, Homicide Severity, and Application of the Death Penalty: A Consideration of the Barnett Scale," 27 *Criminology* 511.

Landwehr, James M., Daryl Pregibon, & Anne C. Shoemaker (1984) Graphical Methods for Assessing Logistic Regression Models," 79 *J. of the American Statistical Association* 61.

Maynard, Douglas W. (1984) *Inside Plea Bargaining: The Language of Negotiation*. New York: Plenum Press.

Neubauer, David W. (1974) *Criminal Justice in Middle America*. Morristown, NJ: General Learning Press.

Paternoster, Raymond (1983) "Race of Victim and Location of Crime: The Decision to Seek the Death Penalty in South Carolina," 74 *J. of Criminal Law & Criminology* 754.

——— (1984) "Prosecutorial Discretion in Requesting the Death Penalty: A Case of Victim-based Racial Discrimination," 18 *Law & Society Rev.* 437.

Paternoster, Raymond, & Annmarie Kazyaka (1988) "Racial Considerations in Capital Punishment: The Failure of Evenhanded Justice," in K. C. Haas & J. A. Inciardi, eds., *Challenging Capital Punishment: Legal and Social Science Approaches*. Newbury Park, CA: Sage Publications.

Radelet, Michael L., & Glenn L. Pierce (1985) "Race and Prosecutorial Discretion in Homicide Cases," 19 *Law & Society Rev.* 587.

Rosett, Arthur, & Donald R. Cressey (1976) *Justice by Consent: Plea Bargains in the American Courthouse*. New York: Lippincott.

Sanders, Andrew (1987) "Constructing the Case for the Prosecution," 14 *J. of Law & Society* 229.
Stanko, Elizabeth A. (1981) "The Arrest versus the Case," 9 *Urban Life* 395.
Tversky, Amos, & Daniel Kahneman (1974) "Judgement under Uncertainty: Heuristics and Biases," 185 *Science* 1124.
U.S. General Accounting Office (1990) *Death Penalty Sentencing: Research Indicates Patterns of Racial Disparities*. Washington, DC: GAO.

Cases Cited

Furman v. Georgia, 408 U.S. 238 (1972).
Godfrey v. Georgia, 446 U.S. 420 (1980).
Gregg v. Georgia, 428 U.S. 153 (1976).
Maynard v. Cartwright, 486 U.S. 356 (1988).
McCleskey v. Kemp, 481 U.S. 279 (1987).
Proffit v. Florida, 428 U.S. 242 (1976).
Woodson v. North Carolina, 428 U.S. 280 (1976).