

# **Model Uncertainty and the Deterrent Effect of Capital Punishment**

**Ethan Cohen-Cole<sup>1</sup>**  
**Steven Durlauf**  
**Jeffrey Fagan**  
**Daniel Nagin**

**February 28, 2007**

## **Abstract**

The reintroduction of capital punishment after the end of the Supreme Court moratorium has permitted researchers to employ state level heterogeneity in the use of capital punishment to study deterrent effects. However, no scholarly consensus exists as to their magnitude. A key reason this has occurred is that the use of alternative models across studies produces differing estimates of the deterrent effect. Because differences across models are not well motivated by theory, the deterrence literature is plagued by model uncertainty. We argue that the analysis of deterrent effects should explicitly recognize the presence of model uncertainty in drawing inferences. We describe methods for addressing model uncertainty and apply them to understand the disparate findings between two major studies in the deterrence literature, finding that evidence of deterrent effects appears, while not nonexistent, is weak.

---

<sup>1</sup>Cohen-Cole: Federal Reserve Bank of Boston, 600 Atlantic Avenue, Boston, MA ethan.cohen-cole@bos.frb.org; Durlauf (corresponding author): Department of Economics, University of Wisconsin. 1180 Observatory Drive, Madison WI, 53706-1393. sdurlauf@ssc.wisc.edu; Fagan: Columbia University School of Law; 435 West 116th Street Room 634, Box D-18 New York NY 10027; Nagin: H. John Heinz III School of Public Policy & Management; Carnegie Mellon University 5000 Forbes Avenue. Pittsburgh, PA 15213-3890. The Department of Justice's National Institute of Justice has provided financial assistance. We are grateful for research assistance provided by Jonathon Larson. Durlauf thanks the National Science Foundation for financial support. The views expressed in this paper are solely those of the authors and do not reflect official positions of the Federal Reserve Bank of Boston or the Federal Reserve System.

## 1. Introduction

This paper explores the strength of evidence on the deterrent effect of capital punishment with specific attention to recent studies of this question. Its goals are both methodological and substantive. In terms of methodology, we argue that an important difficulty with studies of capital punishment is the failure to address systematically model uncertainty; we offer suggestions on how model uncertainty can be addressed. In terms of substance, we perform an exercise that show how these methods can be used to address differences between studies. Our results help to explicate a significant disagreement in the empirical literature.

The new wave of research on capital punishment and deterrence is based on the data that have become available following the reintroduction of capital punishment in different states beginning in 1976. Not all states reintroduced when the Constitutional barrier was lifted: they acted at different times to restore capital punishment and used it at widely varying rates. The resulting natural variation in execution rates across states and time forms the empirical basis for these studies.

This new body of work has failed to produce a consensus on whether deterrent effects are present. Dezhbakhsh, Rubin, and Shepherd (2003, 2006) and Mocan and Gittings (2001) find strong deterrent effects from the death penalty. These claims have been challenged by Donohue and Wolfers (2005), Berk (2005), and Fagan (2006), who argue that the evidence that has been adduced in favor of strong deterrence effects is fragile, in that they may be reversed by small changes in model specification. Other studies have argued that more substantive differences in the formulation of the deterrence mechanism lead to different results. Katz, Levitt, and Shustorovich (2001), focusing on the fact that executions are relatively infrequent, argue that prison mortality rates represent a deterrent whereas capital punishment does not. Other studies find that deterrence effects are heterogeneous, so that important properties are masked by imposing a single measure on the statistical analysis. Shepherd (2005) draws mixed conclusions, suggesting that capital punishment will raise murder rates when the number of executions is small, producing what she calls a brutalization effect. However, the brutalization effect is dominated by the deterrence effect when the number of executions exceeds some empirically identified threshold. Hjalmarsson (2006) explores whether executions have short run local

deterrence effects by studying the city level effects using daily frequency data; focusing on Texas, she finds little evidence of deterrence.

The different empirical studies of deterrence have much in common. They are based on a common choice-theoretic version of criminal behavior advanced by Gary Becker and implemented in the capital punishment context by Isaac Ehrlich (1975, 1977). All the studies reflect the common idea that criminal behavior reflects a purposeful calculation of its benefits and costs. For the study of murder and capital punishment, a choice-theoretic model leads different researchers to employ qualitatively similar factors in understanding how individuals assess benefits and costs; it is not the case, for example, that one study assumes that the propensity to commit a murder is determined by sociological and cultural factors whereas another does not. The different studies are also similar in that they all employ aggregate observational data on murder, punishments risks, and rich sets of covariates to evaluate deterrent effects. Despite these communalities, estimates of deterrence effects vary dramatically and often are contradictory.

Why should studies using relatively similar conceptualizations of individual behavior and similar degrees of data aggregation produce disparate conclusions? A fundamental problem that underlies the disparate findings on the deterrent effect of death sentencing is that individual studies reflect specific assumptions about the appropriate data, control variables, functional form specification, etc. on the part of the researcher. As a result, two researchers—each of whom has developed conceptually a reasonable and potentially correct “model” (by which we mean a collection of assumptions)—can reach opposing conclusions. Typically, these differences in assumptions across models cannot be resolved by appeals to theory or to widely-endorsed statistical practice, because *a priori* each may be argued to be sensible in, at least, some circumstances. *A posteriori*, of course, evidentiary support may favor one model over another. But even if the evidentiary support is quite lopsided, it is not the case that one naturally regards the probability that one set is true and the other false as either 100% or 0%.

Our goal in this paper is to provide a constructive approach to addressing the model uncertainty that is found in the capital punishment literature. As we see it, the objective of deterrence studies is not to identify a best model of the murder process, but to communicate the information embodied in a data set on deterrence *per se*. How might this be done in practice? If the objective of the exercise is to communicate a single estimate of the deterrence effect of an

execution (and an associated measure of the uncertainty of the estimate), then this should be done with explicit recognition of the model uncertainty present in the analysis. This leads us to employ model averaging methods.

In model averaging, the researcher treats the “true” model of a phenomenon as unknown and proceeds through several stages to produce a probability that each of the candidate models is the “true” one. The first step consists of formulation of a space of candidate models. This step involves judgment; there is no algorithm for determining what models should be considered. Second, one determines what information is available in each element of the space. At this stage, it is necessary to take a stand on whether one wants to produce frequentist estimates or Bayesian posterior probabilities of parameters. Third, the information available in the individual models is averaged, accounting for probability that each of the models is, in fact, the true one. These probabilities need to be constructed regardless of whether one is a frequentist or Bayesian.

The model averaging approach was originally suggested in Leamer (1978) but has only recently reappeared in the statistics literature, where Draper (1995) provided a general conceptual argument in favor of model averaging. Also, work by Adrian Raftery (e.g. Raftery, Madigan, and Hoeting (1997)) that has been fundamental in making the approach operational. Model averaging has appeared in a number of social science settings, notably economics, examples include Brock and Durlauf (2001), Doppelhofer, Miller, and Sala-i-Martin (2004), Fernandez, Ley and Steel (2001) in studying economic growth determinants, and Brock, Durlauf, and West (2003,2006) and Levin and Williams (2001) for monetary policy evaluation.

Interestingly, the first wave of capital punishment/deterrence findings was also criticized for failing to account for model uncertainty. Specifically, Ehrlich’s (1975, 1977) results were challenged by Baldus and Cole (1975), Bowers and Pierce (1975), Klein, Forst, and Filatov (1978), Leamer (1983), McManus (1985) and Passell and Taylor (1977) on the grounds of fragility. Each of these critiques used methods to evaluate whether the results found were sensitive to the particular assumptions used in setting up the econometric study, so that different assumptions could succeed or fail to produce evidence of deterrence. A 1978 National Research Council report reached the general conclusion that “available studies of [of capital punishment] provide no useful evidence on the deterrent evidence of capital punishment” (Blumstein, Cohen, and Nagin (1978), pg. 9.)

Relative to our approach, Leamer's (1983) method is of particular interest. It represents an effort to systematically evaluate the interplay of model specification and deterrence findings. By contrast, the bulk of studies responding to Ehrlich aimed at identifying alternative specifications to Ehrlich's that suggested no deterrence. Extreme bounds analysis was used to argue that Ehrlich's findings were not robust to model choice. Extreme bounds analysis treats a parameter estimate as fragile if its sign flips across model specifications. Our approach avoids using this method which, as argued in Brock, Durlauf and West (2003), amounts to a very special view of how a policymaker should assess evidence; specifically the approach assumes that the policymaker possesses preferences such that if a policy may be counterproductive under any of the models that may characterize the data, then the policy should not be adopted. This means that the relative evidentiary support for different models is ignored in the assessment. Ehrlich and Liu (1997) show that a number of propositions that receive essentially unanimous assent among economists as valid would be rejected using extreme bounds analysis on a sensible data set and group of candidate models. Our approach should, therefore, be understood as incorporating Leamer's fundamental insight but extending it in different directions.

In exploring the role of model uncertainty in deterrence regressions, we are able to provide some adjudication of the disparate results in the literature. In particular, we will provide a variety of ways to understand the relationship between the different conclusions drawn by Dezhbakhsh, Rubin and Shepherd (2003) and Donohue and Wolfers (2005). In doing this, we will conclude that the evidentiary support for a deterrence effect, in the data set under study and the model space spanned by these two papers, is relatively weak.

Given the politically charged nature of any claim concerning capital punishment and deterrence, it is important to state up explicitly what one *can* and *cannot* take from our analysis when considering capital punishment as a public policy. Our findings demonstrate that evidence for deterrence is weak in the context of a major data set and set of possible models of the murder process. Hence, in our judgment, claims of a strong deterrent made on the basis of this data and elements of the model set that we study, cannot be sustained and thus should not so be used to bolster a case for capital punishment. However, nothing in our analysis necessarily speaks to the question of the deterrent effects of capital punishment regimes different from that which has been historically observed. Our analysis can only be interpreted as providing evidence on capital punishment/murder patterns as occurred under the particular policy regime that has existed in the

United States since the resumption of capital punishments in 1976 following the U.S. Supreme Court’s decision in *Gregg v. Georgia* (1976).<sup>2</sup> We also do not deal with broader issues of whether regressions of the type we study can be interpreted as providing causal inferences.<sup>3</sup>

Section 2 of this paper outlines our methodology. Section 3 describes the relationship between Dezhbakhsh, Rubin and Shepherd and Donohue and Wolfers, which will form the basis of our empirical analysis. Section 4 discusses implementation issues. Section 5 reports results. Summary and conclusions appear in Section 6.

## 2. Model uncertainty and model averaging: general principles

We follow the discussion in Brock, Durlauf and West (2003) to describe our statistical framework. Let  $\delta$  denote the measure of the deterrent effect of capital punishment. A typical capital punishment paper is designed to produce statements about this measure conditional on a data set  $D$  and a statistical specification, i.e. a given model  $m$ . An empirical paper will rarely report a set of statements about  $\delta$  given a single model, and so one finds reports of estimates of  $\delta$  for some range of alternative specifications to  $m$ . Examples of such alternatives include different choices of control variables, or choices of functional form. So, in this sense, empirical studies typically recognize the presence of model uncertainty. However, they fail to address it in a systematic fashion. Explorations of the robustness of particular findings are made in an *ad hoc* way and in a manner in which the model uncertainty is “local” to the baseline model  $m$ , i.e. the deviations from the baseline are usually modest.

How might model uncertainty be treated in a systematic fashion? Relative to our description of the “standard” empirical exercise, we argue that evidence on  $\delta$  should be reported based upon a model space  $M$  that is constructed to span plausible alternative models. In other words, a researcher needs to explicitly consider what aspects of his model are uncertain, and treat different resolutions of this uncertainty as candidate models. Information about the deterrent

---

<sup>2</sup> See *Gregg vs Georgia*, 428 US 153 (1976).

<sup>3</sup> See Fagan (2006) for discussion. While such issues are obviously exceptionally important, our focus is solely on the model uncertainty question, applied to a statistical framework that has in fact been used for causal claims.

effect should not be based on the assumption that one or a small, arbitrarily chosen subset of these models, are the only ones that should be considered.

How should one describe evidence on a phenomenon such as deterrence when model uncertainty is present? Some intuition to our approach may be derived from considering the question of how to handle disagreements about heterogeneity in the objects studied in a data set.<sup>4</sup> One of the sources of the different findings in Dezhbakhsh, Rubin and Shepherd and Donohue and Wolfers is the choice of data to use. Donohue and Wolfers argue that excluding a single state, such as California or Texas, from the Dezhbakhsh, Rubin and Shepherd data is a major source of the difference in findings (see, also, Berk (2005)). One can think of this disagreement as reflecting a simple form of model uncertainty in that the model space has only two elements: a set of data that includes California and one that does not. How do we propose adjudicating the disagreement? We argue that one should average the estimates from the two studies by taking a weighted average of the results from each, where the weights are model probabilities.<sup>5</sup> Dezhbakhsh, Rubin and Shepherd can be interpreted as placing a prior probability of 1 on the model with data that includes California whereas a researcher using Donohue and Wolfers would place a prior probability of 1 on the model without California. (To be clear, Donohue and Wolfers themselves do not endorse any particular model; rather they use it to illustrate the fragility of the claims in Dezhbakhsh, Rubin and Shepherd.) Our approach recognizes that each model has information that is useful to a researcher.

More formally, the structure of model averaging may be understood as follows. Suppose one wishes to produce an estimate of some object of interest,  $\delta$ , which measures the effects of a policy. In the context of the capital punishment literature,  $\delta$  tends to be the coefficient on the execution variable in some deterrence regression. Conventional statistical methods may be thought of as calculating an estimate that is model-specific,  $\hat{\delta}_m$ . In the model averaging approach, one attempts to eliminate conditioning on a specific model. To do this, one specifies a set or space of possible models  $M$ . The true model is, of course, unknown, so from the perspective of the researcher, each model will have some probability of being true. These

---

<sup>4</sup>These types of disagreements can be formalized using the probabilistic notion of exchangeability, see Brock and Durlauf (2001) for discussion.

<sup>5</sup>This can be thought of simply as the conditional probability that a given model describes the data. We discuss this at greater length below.

probabilities depend on the relative goodness of fit of the different models given available data  $D$  as well as the prior beliefs of the researcher (something we discuss below); hence each model is associated with a posterior probability:  $\mu(m|D)$ . These posterior probabilities allow us to average the model-specific estimates to produce an estimate that accounts for the model uncertainty,

$$\hat{\delta}_M = \sum_m \mu(m|D) \hat{\delta}_m . \quad (1)$$

An associated variance estimate (due to Leamer (1978)) is

$$\text{var}(\hat{\delta}) = \sum_{m \in M} \mu(m|D) \text{var}(\hat{\delta}_m) + \sum_{m \in M} \mu(m|D) (\hat{\delta} - \hat{\delta}_m)^2 . \quad (2)$$

The estimate  $\hat{\delta}_M$  thus accounts for the information contained in each specific model about  $\delta$  and weights this information according to the likelihood the model is the correct one. Brock, Durlauf, and West (2003) argue that the strategy of constructing estimates that are not model-dependent is the appropriate one when the objective of the statistical exercise is to evaluate alternative policy questions such as whether to implement capital punishment in a state. Notice that this approach does not identify the “best” model; instead, it focuses entirely on estimating the effect of the policy, i.e. the parameter  $\delta$ .

The variance formula (2) is interesting as it illustrates how model uncertainty affects the overall uncertainty one should associate with given parameter estimates. The variance of  $\hat{\delta}_M$  consists of two separate parts. The first,  $\sum_{m \in M} \mu(m|D) \text{var}(\hat{\delta}_m)$ , is a weighted average of the variances of the estimates of  $\delta$  for each model and has the same form as the model average estimate of the parameter itself, i.e. (1). The second term  $\sum_{m \in M} \mu(m|D) (\hat{\delta} - \hat{\delta}_m)^2$  does not have analog in (1). It reflects the variance of the parameter estimates across the models in  $M$ ; this variance is produced by the fact that the models are themselves different. This term is not determined by the model-specific variance estimates and thus captures how model uncertainty

increases the variance associated with a parameter estimate relative to conventional calculations. To see why this second term is interesting, suppose that  $\text{var}(\hat{\delta}_m) = 0 \quad \forall m$ , so that conditional on each model, there is no uncertainty about the parameter. While the component  $\sum_{m \in M} \mu(m|D) \text{var}(\hat{\delta}_m)$  will therefore equal 0, it would of course be silly to conclude that the overall variance of the parameter estimate is 0, so long as there is any variation in  $\hat{\delta}_m$ . More generally, the cross-model variation in  $\hat{\delta}_M$  is a distinct source of uncertainty (as measured by the variance) that exists with respect to  $\delta$ .

Notice that averaging across models means that a key role is played by the posterior model probabilities. Using Bayes' rule, the posterior probability may be rewritten as

$$\mu(m|D) = \frac{\mu(D|m)\mu(m)}{\mu(D)} \propto \mu(m)\mu(D|m). \quad (3)$$

The calculation of posterior model probabilities thus depends on two terms. The first,  $\mu(m)$ , is the prior probability assigned to model  $m$ . Computing posterior model probabilities requires specifying prior beliefs on the probabilities of the elements of the model space  $M$ . It is common in the model averaging literature to assume that all models in  $M$  have equal prior probability. While this assumption may be criticized, (Brock, Durlauf, and West (2003), Doppelhofer and Weeks (2006)), for this context the assumption seems reasonable, and we follow it in our empirical implementation. The second term,  $\mu(D|m)$ , is the probability of data given a model. This term ensures that models with greater evidentiary support receive greater weight in evaluating  $\delta$ . An important difference with standard empirical work is that models with relatively weak evidentiary support are not ignored, even if they are downweighted. This represents a difference from most empirical work, which concentrates on first selecting a model, and second drawing inferences based on a parameter within the model. Model selection amounts to assigning a weight of 1 to a particular model given superior empirical performance to others. This exaggerates the empirical salience of the model and thus can lead to inappropriate inferences.

### **3. Dezhbakhsh, Rubin and Shepherd versus Donohue and Wolfers**

One of the most prominent papers arguing for the presence of a deterrent effect of capital punishment is the (2003) study by Dezhbakhsh, Rubin and Shepherd. This study is based on county-level data for the post-moratorium period (1977-1996); at the time it arguably represented the most detailed and disaggregate dataset to have been used to study deterrence and compares favorably with other data sets that have subsequently been used.

The Dezhbakhsh, Rubin and Shepherd model is standard from the perspective of the choice-theoretic model of crime. In the model, the murder rate is a function of three principal deterrence variables: the probability of arrest, the probability of receiving a death sentence conditional on being arrested, and the probability of being executed conditional on receiving a death sentence. The model includes controls for related crime variables including the aggravated assault rate and the robbery rate. Demographic variables include information on population subsamples whose population shares may be correlated with higher levels of crime: the population proportion of 10-19 year olds and 20-29 year olds, percentages of blacks, percentages of non-black minorities, population density and the male population share. Income variables include real per capita income, real per capita income maintenance payments, and real per capita unemployment insurance payments. Finally, the specification includes the percentage of NRA members. These control variables are proxies for heterogeneity in murder rates across demographic groups, the opportunity cost of crime (proxied by various economic measures), as well as access to weapons. While one can naturally question the mapping between the empirical proxies and the actual determinants of murder, the variable choices reflect the constraints imposed by data availability and are in fact quite conventional. Formally, the Dezhbakhsh, Rubin and Shepherd murder rate regression is

$$\begin{aligned}
\frac{Murders_{c,s,t}}{pop_{c,s,t}} = & \delta_0 + \delta_1 \frac{HomicideArrests_{c,s,t}}{Murders_{c,s,t}} + \delta_2 \frac{DeathSentences_{s,t}}{HomicideArrests_{s,t-2}} + \delta_3 \frac{Executions_{s,t}}{DeathSentences_{s,t-6}} \\
& + \gamma_1 \frac{Assaults_{c,s,t}}{Population_{c,s,t}} + \gamma_2 \frac{Robberies_{c,s,t}}{Population_{c,s,t}} + \gamma_3 Demographics_{c,s,t} + \\
& + \gamma_5 economy_{c,s,t} + \gamma_6 \frac{NRAMembers_{s,t}}{population_{s,t}} + \sum_c \gamma_{7,t} county_c + \sum_t \gamma_{8,t} time_t + \eta_{s,t} + \varepsilon_{c,s,t}
\end{aligned} \tag{4}$$

The overall deterrence effect of capital punishment can be evaluated by viewing the parameters  $\delta_1$ ,  $\delta_2$ , and  $\delta_3$ . Estimates from Dezhbakhsh, Rubin and Shepherd and Donohue and Wolfers use  $\delta_3$  exclusively to determine the number of lives saved or lost from executions; we follow this method. Since each of the variables associated with these parameters is endogenous, these parameters are estimated using instrumental variables based on the following first stage regressions<sup>6</sup>

$$\frac{HomicideArrests_{c,s,t}}{Murders_{c,s,t}} = \psi_0 + \psi_1 \frac{Murders_{c,s,t}}{Pop_{c,s,t}} + \psi_2 PolicePayroll_{s,t} + \sum_t \psi_{3,t} Time_t + \varepsilon'_{c,s,t} \tag{5}$$

$$\begin{aligned}
\frac{DeathSentences_{s,t}}{HomicideArrests_{s,t}} = & \theta_0 + \theta_1 \frac{Murders_{c,s,t}}{pop_{c,s,t}} + \theta_2 JudicialExpense_{s,t} + \theta_3 PartisanInfluence_{s,t} \\
& + \theta_4 Admissions_{s,t} + \sum_t \theta_{5,t} Time_t + \varepsilon''_{c,s,t}
\end{aligned} \tag{6}$$

and

$$\begin{aligned}
\frac{Executions_{s,t}}{DeathSentences_{s,t}} = & \phi_0 + \phi_1 \frac{Murders_{c,s,t}}{Pop_{c,s,t}} + \phi_2 JudicialExpense_{s,t} + \phi_3 PartisanInfluence_{s,t} \\
& + \sum_t \phi_{4,t} Time_t + \varepsilon'''_{c,s,t}
\end{aligned} \tag{7}$$

---

<sup>6</sup>While we follow Dezhbakhsh, Rubin and Shepherd in instrumental variable choice, there are good reasons to challenge their validity of these instruments as nicely described in Donohue and Wolfers (2006). Rubin's (2006) reply to this criticism is nonresponsive to the substance as he does nothing more than say 1) finding valid instruments is hard and 2) suggesting the force of the Donohue and Wolfers criticism is weakened by the fact they did not make it earlier!

The variable  $Pop_{c,s,t}$  indicates the population in county  $c$ , state  $s$ , and time  $t$ , divided by 100,000.  $PartisanInfluence$  is measured by the Republican presidential candidate's vote share in the most recent election, and  $Admissions$  is the prison admission rate. Note that some of the key variables are estimated at the state level (the subscript  $c$  is omitted in these cases).<sup>7</sup> Additional information is available in the original study.

Dezhbakhsh, Rubin and Shepherd (p 362-363) present the results from this given six different versions of the variable  $\frac{Executions}{DeathSentences}$ . They consider 3 different measures of execution probabilities in the six columns of their Tables 3 and 4 are as follows:

$$\text{Columns 1 and 4: } \frac{Executions_{s,t}}{DeathSentences_{s,t-6}} \quad (8)$$

$$\text{Columns 2 and 5: } \frac{Executions_{s,t+6}}{DeathSentences_{s,t}} \quad (9)$$

$$\text{Columns 3 and 6: } \frac{\sum_{t=-3}^3 Executions_{s,t}}{\sum_{t=-9}^{-4} DeathSentences_{s,t}} \quad (10)$$

Columns 1-3 omit observations in which there are no death sentences. Columns 4-6 use a method in which the probability is based on the most recent year which had a death sentence. Table 7 (p 824) in Donohue and Wolfers replicates the original results as well as reports their own findings for different specifications. While both papers provide a wide variety of other analyses, we focus on this Table from Donohue and Wolfers both because it provides a useful case to illustrate our primary arguments and because it captures the main differences in the findings of the studies.

Donohue and Wolfers (2005) challenge the Dezhbakhsh, Rubin and Shepherd findings on the grounds of fragility. Specifically, they show that three modifications to the Dezhbakhsh,

---

<sup>7</sup>Dezhbakhsh, Rubin and Shepherd use a combination of county and state effects to predict county-level murder rates. We will not discuss the merits (or difficulties) of this type of estimation strategy other than to comment that it will not impact the model averaging exercise that we are conducting.

Rubin and Shepherd construction can strongly affect findings of a strong deterrence effect to capital punishment. The changes are 1) the use of a single voting variable instead of six in the first stage regressions, 2) omission of Texas from the analysis and 3) omission of California from the analysis. These changes cause the sign of the estimate deterrent effect to reverse. That is, under this alternative specification, each execution is predicted to *increase* the murder rate. Table 1 summarizes the different assumptions in the two papers.

The differences between the Dezhbakhsh, Rubin and Shepherd and Donohue and Wolfers findings illustrate how model uncertainty matters for substantive empirical claims, even when this uncertainty is predicated a common social science theoretical structure. Each of these papers takes a particular stand on choice of instrumental variables and the nature of inter-state comparability in the data under study. At the same time, both studies use the same choice-theoretic approach to criminal behavior; however, the approach does not provide any theoretical guidance on the correct statistical model. Because theory fails to identify the appropriate statistical specification, model averaging is thus a natural way to proceed.

#### **4. Implementation issues**

##### **Data**

With the exception of one variable based on data from the National Rifle Association, the data used in this analysis are publicly available from the FBI's Uniform Crime Reports, the Department of Justice's Bureau of Justice Statistics (BJS), and the Bureau of the Census. Donohue and Wolfers (2005), who successfully replicate the results in Dezhbakhsh, Rubin and Shepherd kindly provided their data to us and made it publicly available at <http://bpp.wharton.upenn.edu/jwolfers/DeathPenalty.shtml>. All of our analyses are based on this source as Dezhbakhsh, Rubin and Shepherd declined our request to share the original data set used in their study.

The data are a panel of state and county level data covering the time period from 1977-1996. The FBI was the source of information on crime and arrest rates. The BJS was the source on police and judicial expenditure which is used to control for variation in the probability of

being caught and being sentenced, respectively. To account for variation in execution rates, BJS data on the number of executions was used. The BJS was also the source of data on prison populations, prison admittances and number of death sentences. Demographic information including age, sex, race, and geographic size of counties are from the US Bureau of the Census. We further employ the Republican voting share in the prior presidential election as a control for social concern with crime. Economic information includes income, unemployment, income maintenance, and retirement payments are from the Bureau of Economic Analysis. NRA membership rates come directly from the National Rifle Association.

### **Model space construction**

To implement model averaging using available methods, specifically those developed in Raftery, Madigan and Hoeting (1997), we translate the differences between Dezhbakhsh, Rubin and Shepherd and Donohue into differences in the choice of control variables at different stages of the analysis. With respect to the choice of instrumental variables, it is by definition a question of variable inclusion. In order to model heterogeneity between California, Texas and the rest of the United States as a matter of variable inclusion, we proceed differently from Donohue and Wolfers. Rather than omit these states from the data under study, we construct variables that are the products of the deterrence variables and a dummy for California, and corresponding variables that are the product of the deterrent variables with a dummy for Texas. Note that this approach to dealing with California and Texas is substantively different from Donohue and Wolfers in that the California and Texas data are retained and will affect all model parameter coefficients. We include these new variables in our study in two ways. Our model averaging exercise introduces 6 possible independent variables (2 variables each corresponding to the 3 key deterrence variables) to augment the original Dezhbakhsh, Rubin and Shepherd model.<sup>8</sup> The consequent

---

<sup>8</sup>Our procedure allows the possibility that one of the deterrence variables, differs between a state (e.g. California) and the rest of the states, whereas the others are not allowed to so vary. This might seem odd in that one would expect heterogeneity to apply to all the deterrence variables or none. We therefore also considered a more 'limited' model space in which only 2 *sets* of variables enter the model space. The 2 variable sets are those that correspond to the 3 deterrence variables interacted with one of the state dummies. Thus, in this version, the space effectively includes 2 new 'elements', each of which consists of 3 variables. As the analyses with this

model space is thus the set of combinations of the instrumental variables used in Dezhbakhsh, Rubin and Shepherd and Donohue and Wolfers, in addition to the variables constructed here.

We treat each choice of model as choice of first and second stage regressions. Specifically, each model represents a choice of instrumental variables for the equations (5)-(7) and a choice of control variables for the second equation (4). Different instrumental variables choices are generated by different sets of voting variables. Different second-stage regressions are determined by the sorts of heterogeneity between California, Texas and the rest of the United States that is allowed; this is done via the particular set of deterrence variables interacted with state dummies that is included. Our model space thus consists of 384 different specifications. Specifically, we include all combinations of Texas and California dummies interacted with the three deterrence variables. This produces 6 variables and 64 possible models ( $2^6$  combinations). For each of these, we consider the 6 possible deterrence definitions from DRS – producing our 384 total.

### Model weight calculation

Model weights are according to the selection of variables and model fit in the second stage regressions. Our calculation replicates Raftery (1995) so that

$$p(M | D) \approx \exp\left(-\frac{1}{2} BIC_k\right) / \sum_{l=1}^k \exp\left(-\frac{1}{2} BIC_l\right) \quad (11)$$

where  $BIC$  is defined as

$$BIC = n \log(1 - R_p^2) + p \log n \quad (12)$$

---

limited model space did not vary in interesting ways from the analysis with the larger model space, we do not report a separate analysis for reasons of space.

and  $p$  is the number of regressors, and  $R_p^2$  is the generalized measure of goodness of fit for instrumental variables regressions proposed by Pesaran and Smith (1994).

Our model weights are chosen to provide a simple way of aggregating information across models. As such, our exercise is meant to illustrate the information on capital punishment in a particular data set and give an indication of how model specification affects that information. There is, as far as we know, no formal analyses of model averaging for instrumental variables contexts that would provide a formal justification of the weights in terms of a formal Bayesian procedure, for example one with diffuse coefficient priors within models, as exists for ordinary least squares contexts. That said, the Pesaran and Smith goodness of fit measures do provide a consistent way of comparing choices of instrumental variables across models, and the *BIC* does provide a penalty for model complexity, so we believe the weights are sensible. We also note that we will report some properties of the model-specific estimates that do not use model weights; these produce qualitatively similar conclusions.

In interpreting our results, it is important to recognize that the distribution of the model averaged parameter estimates and associated standard errors cannot be formally equated with the sorts of distributions conventionally used to conduct hypothesis testing. Our focus is instead on whether standard error estimates are large compared to parameter estimates. Our rule of thumb in the discussion is that a standard error is small when it is less than  $\frac{1}{2}$  the value of the associated parameter estimate, large otherwise. This roughly corresponds to the treatment of  $t$ -statistics greater than 2 as statistically significant.

## 5. Results

Recall that Dezhbakhsh, Rubin and Shepherd report deterrence findings for six different constructions (eq. 8-10) of the probability of execution given a death sentence. Table 2 reports model averaged deterrent effects for each of these six constructions. As the Table indicates, there is no consistent message about the strength of the deterrent effect from executions. For two of the six categories (columns 4 and 6), the estimate of the net lives saved is positive and approximately 1.5 to twice as large as the associated standard error. For two of the categories (columns 2 and 5) the estimated number of net lives saved is positive, with a standard error of

approximately the same size or larger than the estimate. Finally, for two categories the net lives saved estimate is negative, but with large associated standard errors. This inconsistency contrasts with the Dezhbakhsh, Rubin and Shepherd finding of a statistically significant positive net lives saved estimate for each category, with standard errors consistently less than half as large as the estimates. The findings do not mirror the findings of Donohue and Wolfers in the sense that for some categories, they found that alternative assumptions lead to a statistically significant association of additional murders for each execution. But this does not contradict the claim of Donohue and Wolfers that Dezhbakhsh, Rubin and Shepherd's findings are model-specific. Our averaging exercise reinforces the Donohue and Wolfers claim. It demonstrates that their conclusion is not an artifact of their having reported a particularly unfavorable alternative specification relative to the Dezhbakhsh, Rubin and Shepherd baseline.

Some additional evidence of the lack of firm deterrence evidence may be obtained from a consideration of the three deterrence parameters  $\delta_1$  (arrest probability),  $\delta_2$  (death sentence probability given arrest), and  $\delta_3$  (probability of execution given death sentence). In principle, the choice-theoretic logic underlying the murder rate regressions implies that each of parameters is negative. Our model-averaged estimates, however, do not show any consistent finding of this type across categories. The signs of the coefficients vary across categories for all three deterrence parameters; the coefficient estimates are generally small compared to the standard errors. Interestingly, the one exception to the imprecise coefficient estimates occurs for the arrest probability coefficient in column 1-3, but here the coefficient is positive, which would mean that more arrests lead to more murders. Notice that columns 4 and 6, which were the cases where we found the strongest evidence that an execution saves additional lives, the deterrence coefficients  $\delta_1$  and  $\delta_2$  do not have the expected signs (though each has a large standard error).

How does the information in the various categories combine to produce an overall deterrence estimate? Table 3 integrates the information across the Dezhbakhsh, Rubin and Shepherd categories in order to produce overall estimates of the deterrence variables and the expected number of lives saved. In introducing this additional level of averaging, we use equal weights across the six categories rather than construct a new model space and then computing

posterior model probabilities; this allows the easiest comparisons with Table 2.<sup>9</sup> As the Table indicates, the summary deterrence statistic, expected number of lives saved per execution, is large relative to most studies, 32, but the standard deviation of the estimate is quite high as well, 26. This suggests that the evidentiary support for a deterrence effect is weak when one reduces the different specifications spanned in the two papers down to a single calculation. When one considers individual deterrence coefficients, one again finds little evidence of deterrence effects, with large standard errors for each estimate and a positive sign for the overall estimate of  $\delta_2$  (death sentence likelihood given arrest) in contradiction to what theory predicts.<sup>10</sup> Together, these findings do not provide a clear and coherent view of deterrent effects for murder. So in this sense, the Donohue and Wolfers critique is shown to hold when one systematically considers model uncertainty.

Some additional insights into the sensitivity of deterrent effect estimates to model specification are illustrated in Table 4. In this table we compare the some of the Dezhbakhsh, Rubin and Shepherd and Donohue and Wolfers with the elements of the model space that provide the largest and smallest estimates of the deterrent effects. To be precise, Column 1 of Table 4 reports the smallest number of net lives saved for any specification we have studied, varying across all the models that are used in the Table 3 averages. Column 2 reports the results for the model with the largest number of net lives saved across all specifications. Column 3 reports the model with the largest posterior model probability of all those considered. Column 4 reports the Dezhbakhsh, Rubin and Shepherd specification with the largest net lives saved per execution. This corresponds to column 3 of Dezhbakhsh, Rubin and Shepherd (pages 362-363, Tables 3 and 4). Column 5 provides a simple average of the Donohue and Wolfers findings. As illustrated in Table 1, Donohue and Wolfers provide three variations on the 6 specifications of Dezhbakhsh, Rubin and Shepherd. We average all 18 of these for the result in Column 5. (We average the Donohue and Wolfers results since their focus was on the fragility of Dezhbakhsh, Rubin and Shepherd, not promoting a particular model.)

---

<sup>9</sup>Unequal weighting across categories would have “violated” the DRS and DW decision to treat the different ways of measuring the probability of execution given a death sentence as equally plausible.

<sup>10</sup>One reason for the anomalous sign may be the high reversal rate on death sentences, see Liebman, Fagan and West, *Texas Law Review* (2000) (68% reversal rate on death sentences). Only about one death sentence in 9 survives to the execution stage.

These results help place the differing conclusions of Dezhbakhsh, Rubin and Shepherd and Donohue and Wolfers in context. They suggest that Dezhbakhsh, Rubin and Shepherd's strong claims on deterrence are not the result of data mining *per se*; there is an alternative specification with far greater net lives saved estimates, and the most favorable model among the limited set they considered is close in net lives saved to the model with highest posterior weight of being true.<sup>11</sup> Instead, the message of this comparison is that Dezhbakhsh, Rubin and Shepherd's findings are driven by their having focused on a particular subset of plausible models. However, when one expands the space of possible models to include a larger set of plausible ones, the evidence for deterrence is greatly weakened. The new models contain additional information about deterrence. We would also note that a comparison of Column 1 with Column 5 illustrates that the Donohue and Wolfers analysis did not focus on outlier models in the sense that Column 1 illustrates a far larger increase in murders from deterrence than the Donohue and Wolfers average; Figure 2 below indicates this is also true for the components of the average.

Finally, we move away from summaries of the behavior of deterrent effects across the model space. Figure 1 provides a weighted histogram of the net lives saved for all the models we have considered. This includes models for each of the 6 Dezhbakhsh, Rubin and Shepherd categories. The weights are the posterior model probabilities. Figure 2 provides the same histogram when the models are all assigned equal weights. For both histograms, observations are placed into 'bins' along the range illustrated in the Figures. In the case of the weighted histogram, each observation is given a frequency weight (measured along the vertical axis) according to its posterior probability. The corresponding measure for an unweighted histogram of course weight each model equally. The Figures indicate the locations in the histograms of Dezhbakhsh, Rubin and Shepherd and Donohue and Wolfers models with the largest and smallest (negative) number of net lives saved across those they considered.

---

<sup>11</sup> The model with the greatest posterior includes is that based on the construction of  $\frac{\text{Executions}}{\text{DeathSentences}}$  corresponding to Table 2, Column 5. It include five of the six interacted deterrent variables, the exception is the California dummy interacted with the probability of death sentence given arrest.

The Figures illustrate the substantial heterogeneity that is present in the model-specific estimates of net lives saved. It is useful to note that the posterior probabilities for all the models in Figure 1 which produce a deterrent effect is 0.76; in Figure 2, the corresponding the percentage of models with positive deterrent effects is 0.72. So, in this sense, there is some evidentiary support for claims of deterrence. That said, the histograms reveal substantial bunching near the origin. In comparing the two histograms, the main difference is that there are a set of models in the unweighted histogram which are associated with large net lives saved estimates which are nearly invisible in the weighted histogram; this is a consequence of the fact that the models have very small posterior probabilities.

## 6. Conclusions

The empirical study of deterrence effects is an example of a problem domain where theory, economic or otherwise, does not provide strong guidance on how to construct the statistical model that maps theory to empirical work. This openedness of different theories of the murder rate, to use a phrase of Brock and Durlauf (2001), means that theory cannot be a precise guide to statistical specification; model uncertainty is intrinsic to such studies (cf. Fagan (2006)). Given that the goal of such exercises is the measurement of a particular policy effect, i.e. the relationship between capital punishment and the murder rate, as opposed to the construction of a model of the murder rate *per se*, model averaging methods are a natural way to make empirical claims robust to the details of model specification. The model averaging approach indicates how one can understand and resolve disparate empirical findings. Our application to the analyses of Dezhbakhsh, Rubin and Shepherd and Donohue and Wolfers leads us to support the conclusion that Dezhbakhsh, Rubin and Shepherd claims about strong deterrence effects are an artifice of particular model choices. On the other hand, we do not find evidence in support of the suggestion that the death penalty raises the homicide rate, which one could take from some of the Donohue and Wolfers regressions (but which the authors did not).

The bottom line of our empirical analysis is that the measure of the number of lives saved per execution is large (weighted average estimate of 36) but imprecisely estimated (weighted average standard deviation 26). This may seem frustrating, since it makes clear that our

conclusion is not that there is no deterrent effect present, but rather that inferences on its magnitude are so imprecise to make representation of strong claims impossible and also irresponsible. But this should not be surprising. As noted by Berk (2005), the analysis of capital punishment and deterrence is naturally delimited by the infrequency of executions for all but a handful of states; in particular, uncertainty about the comparability of Texas naturally produces uncertainty about the overall deterrence estimate given the concentration of executions in that state. The deterrence claim may seem somewhat strengthened by the fact that the posterior model probabilities for those specifications that produce a positive estimate of net lives saved is .76, but in isolation this is a weak piece of evidence since it does not reflect the magnitudes of deterrent effects for those models, let alone the magnitudes for those models where the estimate of net lives saved is negative. It is also important to keep in mind that the weakness of the evidentiary support for deterrence that emerges in our exercise emerges in a context that delimits the model uncertainty in many ways. For example, while we address the issue of the comparability of California and Texas to the rest of the country, we do not address broader issues of the comparability of interstate or intercounty data.

The substantial uncertainty we find associated with deterrence estimates is naturally of importance in moving from positive to normative discussions of the death penalty. Figures such as the expected number of lives saved per execution, or the probability that the deterrence effect of capital punishment is positive do not provide, by themselves, provide guidance as to how a policymaker should use these numbers when comparing policy choices. Should a policymaker care if the conditional probability density under a capital punishment regime is sensitive to what seem to be uninteresting assumptions, such as the choice of instrumental variables? Does a policymaker wish to implement capital punishment, knowing that for a set of *a priori* plausible models, the expected number of murders will increase under the policy? We do not propose answers to these questions, but simply observe that these questions are fundamental to the assessment of capital punishment as a public policy. At a minimum, it seems obvious that a policymaker's preferences might incorporate risk aversion when assessing deterrent effects. Further, there are reasons to believe that policymakers might want to treat model uncertainty differently from other types of uncertainty when assessing policy effects. One reason may be that the policymaker's preferences embody ambiguity aversion, which means that the policymaker has a distaste for the least favorable outcome in an uncertain environment beyond

its role in affecting expected value calculations. These types of preferences may be related to efforts to develop non-Bayesian approaches to decisionmaking.<sup>12</sup> For our purposes, they suggest that simply computing expected deterrence effects is inadequate.

Further progress in evaluating deterrence, in our view, therefore requires the consideration of questions of criminal policy evaluation in light of the types of information limitations we have discussed. Such an analysis will, to be credible, necessarily have to deal with policymaker objective functions that incorporate a richer set of factors than simply minimizing the number of murders.<sup>13</sup> The difficulties in evaluating policies in the presence of efficiency and ethical considerations are considered in Durlauf (2006) and Gaus (2006) among other places; the former paper's analysis of racial profiling has parallels to the capital punishment issue. From the perspective of policy evaluation, the conclusion of National Research Council Report (Blumstein, Cohen and Nagin (1978)) thus still seems appropriate:

Our conclusion about the current evidence does not imply that capital punishment should or should not be imposed. The deterrent effectiveness of capital punishment is only one consideration among many in the decision regarding the use of the death penalty – and, in that decision, those other considerations are likely to dominate the inevitable crude estimates of the deterrent effect (p. 9)

---

<sup>12</sup>This type of work has become important in macroeconomics, see Hansen and Sargent (2007) for a brilliant general exposition; Brock, Durlauf and West (2003) and Brock, Durlauf, Nason, and Rondina (2007) explore policy evaluation issues. Hansen and Sargent defend the use of minimax evaluation on the part of a decisionmaker. Other proposals include the use of minimax regret (Brock, Durlauf, Nason, Rondina (2007)). As noted by Hansen and Sargent, for example, findings in the behavioral economics literature suggest that individual preferences exhibit ambiguity aversion with respect to model uncertainty. When applied to a policymaker, this suggests a sensitivity to the least favorable model that is not captured by our model averaging estimates.

<sup>13</sup>Sunstein and Vermeule's analysis, focuses on evaluating capital punishment when it leads to a reduction of the murder rate, and ignores how to think about appropriate levels of risk or ambiguity aversion on the part of the policymaker; they also largely ignore any competing ethical claims on policy choices.

**Table 1: Dezhbakhsh, Rubin and Shepherd / Donohue and Wolfers Differences**

|                          | DRS<br>Baseline | DW version 1 | DW version 2 | DW version 3 |
|--------------------------|-----------------|--------------|--------------|--------------|
| <i>PartisanInfluence</i> | 6               | 1            | 6            | 6            |
| Texas data               | Include         | Include      | Exclude      | Include      |
| California data          | Include         | Include      | Include      | Exclude      |

*PartisanInfluence* is the Republican vote share in the most recent presidential election. Where the number ‘6’ appears indicates the use of six variables, indicating the share of republican votes in each of the six elections in the dataset. The number ‘1’ indicates the use of a single variable with the share of vote in the most *recent* election at the time in question. DW version 1 uses a single voting variable instead of six in the first stage regressions, version 2 omits Texas from the analysis and version 3 omits California.

**Table 2**  
**Model Averaged Deterrent Effects**

The coefficients in this table are estimated by iterating over the six interaction variables specifying interactions between state dummies for TX and CA and the deterrence variables. The remainder of the DRS controls are used as indicated in their paper.

| <i>Dependent Variable: Annual Homicides per 100,000 Residents</i> |                   |                   |                   |                   |                  |                    |
|---|-------------------|-------------------|-------------------|-------------------|------------------|--------------------|
|   | (1)               | (2)               | (3)               | (4)               | (5)              | (6)                |
| Probability of Arrest   | 1.28<br>(0.24)    | 1.30<br>(0.22)    | 1.29<br>(0.24)    | 2.86<br>(8.72)    | -1.38<br>(6.63)  | 3.45<br>(8.77)     |
| Probability of Death Sentence Given Arrest                        | -30.21<br>(20.16) | -20.43<br>(16.24) | -41.16<br>(21.38) | 107.62<br>(92.93) | 84.25<br>(28.23) | 153.94<br>(113.52) |
| Probability of Execution Given Death Sentence                     | 1.51<br>(6.02)    | -2.33<br>(4.74)   | 2.59<br>(4.24)    | -15.99<br>(10.74) | -3.33<br>(3.27)  | -20.11<br>(11.85)  |
| <b>Implied Life-Life Tradeoff</b>                                 |                   |                   |                   |                   |                  |                    |
| Net Lives Saved   | -12.31<br>(42.27) | 19.02<br>(33.29)  | -21.16<br>(29.77) | 130.61<br>(75.44) | 27.22<br>(22.96) | 164.18<br>(83.21)  |

Controls include the assault rate; the robbery rate; real per capita personal income; real per capita unemployment insurance payments; real per capita income maintenance payments; population density; the proportion of the population aged 10-19, 20-29; black, white, or other; male or female; state NRA membership; and Ordinary Least Squares estimation. Standard errors are in parentheses, and \*\*\*, \*\*, and \* denote statistically significant at 1%, 5%, and 10%, respectively. (a) Implied life-life tradeoff reflects net lives saved evaluated for a state with the characteristics of the average death penalty state in 1996. Instrumental variables regressions are used. Endogenous independent variables are shown in panel A. Instruments include state-level police payroll, judicial expenditures, Republican vote shares, and prison admissions. Controls include the assault rate; the robbery rate; real per capita personal income; real per capita unemployment insurance payments; real per capita income maintenance payments; population density; the proportion of the population aged 10-19, 20-29; black, white, or other; male or female; and state NRA membership.

**Table 3**

**Deterrence Effects: Full Averaging**

| Dependent Variable: <i>Annual Homicides per 100,000 Residents</i> |                    |
|---|--------------------|
|   |                    |
|   |                    |
| Probability of Arrest   | -2.026<br>(4.13)   |
| Probability of Death Sentence Given Arrest                        | 10.028<br>(48.74)  |
| Probability of Execution Given Death Sentence                     | -4.740<br>(6.81)   |
|   | Life-Life Tradeoff |
| Net Lives Saved   | 32.97<br>(26.37)   |

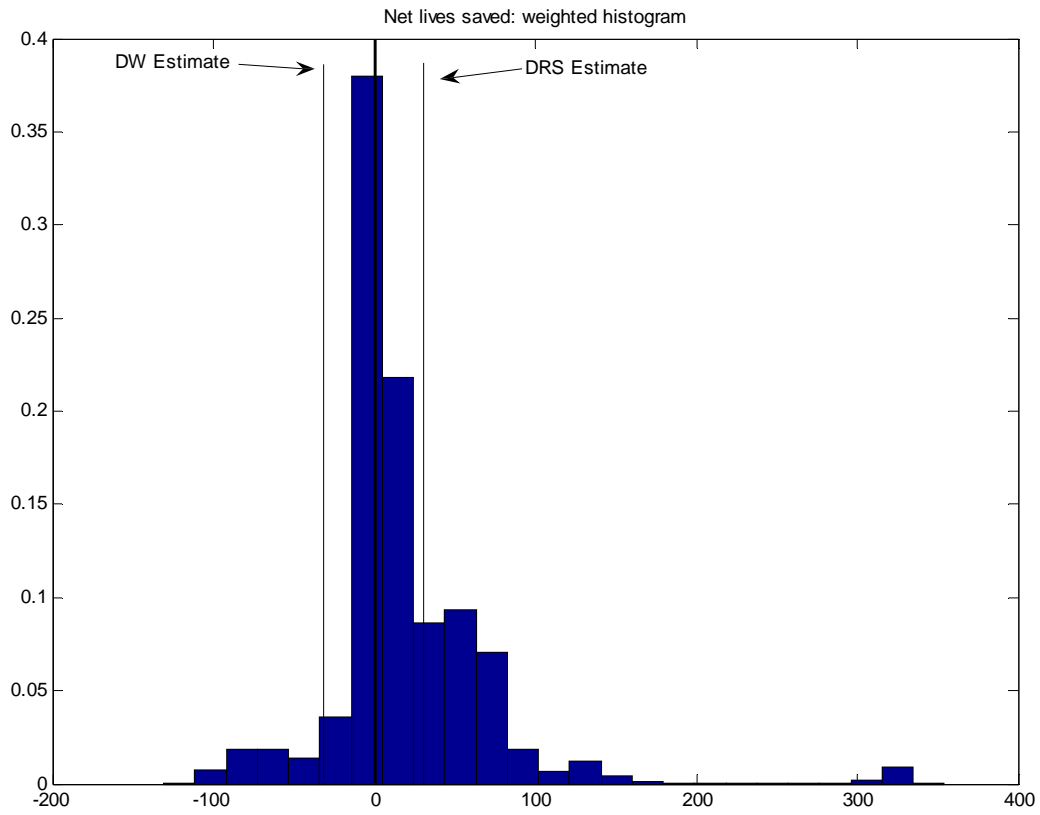
Controls include the assault rate; the robbery rate; real per capita personal income; real per capita unemployment insurance payments; real per capita income maintenance payments; population density; the proportion of the population aged 10-19, 20-29; black, white, or other; male or female; state NRA membership; and Ordinary Least Squares estimation. Standard errors are in parentheses, and \*\*\*, \*\*, and \* denote statistically significant at 1%, 5%, and 10%, respectively. (a) Implied life-life tradeoff reflects net lives saved evaluated for a state with the characteristics of the average death penalty state in 1996. Instrumental variables regressions are used. Endogenous independent variables are shown in panel A. Instruments include state-level police payroll, judicial expenditures, Republican vote shares, and prison admissions. Controls include the assault rate; the robbery rate; real per capita personal income; real per capita unemployment insurance payments; real per capita income maintenance payments; population density; the proportion of the population aged 10-19, 20-29; black, white, or other; male or female; and state NRA membership. Full averaging is the unweighted grand mean taken over the 6 columns in Table 2.

**Table 4 Comparison of Deterrent Effects Estimates**

| Dependent Variable: <i>Annual Homicides per 100,000 Residents</i> |                  |                  |                   |                   |                  |
|---|------------------|------------------|-------------------|-------------------|------------------|
|   | 1                | 2                | 3                 | 4                 | 5                |
|   | Smallest         | Largest          | Largest Posterior | “best”<br>DRS     | Average<br>DW    |
| Probability of Arrest   | 1.68<br>(0.02)   | 19.96<br>(0.58)  | -7.66<br>(0.68)   | -3.33<br>(0.52)   | -4.29<br>(0.54)  |
| Probability of Death Sentence Given Arrest                        | -67.85<br>(7.71) | 393.94<br>(8.88) | 55.62<br>(6.59)   | -32.12<br>(16.22) | -9.03<br>(13.40) |
| Probability of Execution Given Death Sentence                     | 14.84<br>(1.29)  | -44.58<br>(1.27) | -6.32<br>(0.48)   | -7.40<br>(0.72)   | -0.57<br>(0.72)  |
| Implied Life-Life Tradeoff  |                  |                  |                   |                   |                  |
| Net Lives Saved   | -121.2<br>(9.06) | 364.02<br>(8.91) | 51.6<br>(3.37)    | 52.0<br>(5.1)     | 0.08<br>(5.1)    |

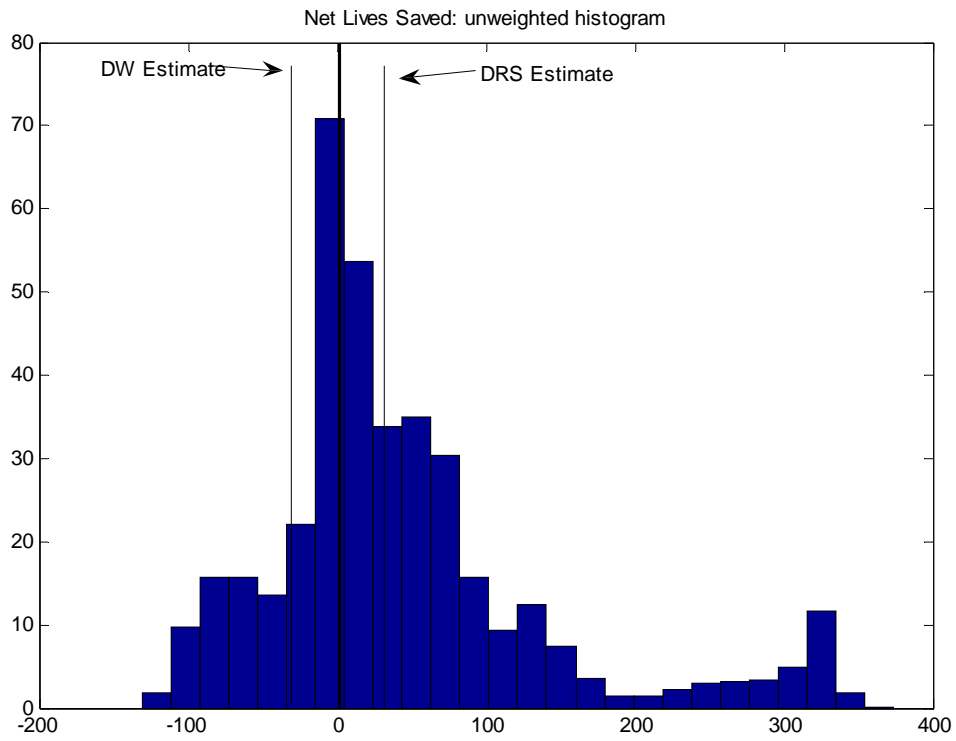
Controls include the assault rate; the robbery rate; real per capita personal income; real per capita unemployment insurance payments; real per capita income maintenance payments; population density; the proportion of the population aged 10-19, 20-29; black, white, or other; male or female; state NRA membership; and Ordinary Least Squares estimation. Standard errors are in parentheses, and \*\*\*, \*\*, and \* denote statistically significant at 1%, 5%, and 10%, respectively. (a) Implied life-life tradeoff reflects net lives saved evaluated for a state with the characteristics of the average death penalty state in 1996. Instrumental variables regressions are used. Instruments include state-level police payroll, judicial expenditures, Republican vote shares, and prison admissions. Controls include the assault rate; the robbery rate; real per capita personal income; real per capita unemployment insurance payments; real per capita income maintenance payments; population density; the proportion of the population aged 10-19, 20-29; black, white, or other; male or female; and state NRA membership. Column 1 corresponds to the smallest number of lives saved, column 2 to the largest, and column 3 to the model with the largest posterior. The DRS results with the largest number of lives saved is shown in column 4 and the average of the DW results shown in column 5. The smallest model includes all control variables and all CA and TX interaction dummies save the CA dummy interacted with the probability of arrest, and uses the sixth definition above. The largest includes control variables and only the TX interaction with the execution variable, and uses the first definition above. The model with the greatest posterior includes all variables from the full specification except the CA dummy interacted with the probability of death sentence given arrest, and uses the fifth definition above.

**Figure 1**



This figure is a weighted histogram of the net lives saved for all the models we have considered, including models for each of the Dezhbakhsh, Rubin and Shepherd categories. The weights are the posterior model probabilities. The DRS and DW lines correspond to the individual model from each with the largest and smallest number of lives saved respectively.

**Figure 2**



This figure provides the same histogram as figure 1 when the models are all assigned equal weights. The DRS and DW lines correspond to the individual model from each with the largest and smallest number of lives saved respectively.

## Bibliography

Baldus, D. and J. Cole, (1975), "A Comparison of the Work of Thorsten Sellin and Isaac Ehrlich on the Deterrent Effect of Capital Punishment," *Yale Law Journal*, 85, 170-184.

Berk, R., (2005), "New Claims about Executions and General Deterrence: Deja Vu All Over Again?," *Journal of Empirical Legal Studies*, 2, 2, 303-330.

Blumstein, A., J. Cohen, J., D. Nagin, D., (1978), *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*, Washington, D.C.: National Academy of Sciences.

Bowers, W. and G. Pierce, (1975), "The Illusion of Deterrence in Isaac Ehrlich's Research on Capital Punishment," *Yale Law Journal*, 85, 187-208.

Brock, W., and S. Durlauf, (2001), "Growth Economics and Reality," *World Bank Economic Review*, 15, 229-272.

Brock, W., S. Durlauf, J. Nason, and G. Rondina, (2007), "Is the Taylor Rule Enough?," mimeo University of Wisconsin

Brock, W., S. Durlauf, and K. West, (2003), "Policy Evaluation in Uncertain Economic Environments," *Brookings Papers on Economic Activity*, 2003, 1, 235-322.

Dezhbakhsh, H., P. Rubin, and J. Shepard, (2003), "Does Capital Punishment Have a Deterrent Effect? New Evidence from Post-Moratorium Panel Data," *American Law and Economics Review*, 5, 2, 344-76.

Dezhbakhsh, H. and J. Shepard, (2006), "The Deterrent Effect of Capital Punishment: Evidence from a 'Judicial Experiment'," *Economic Inquiry*, 44, 3, 512-535.

Donohue, J. and J. Wolfers, (2005), "Uses and Abuses of Empirical Evidence in the Death Penalty Debate," *Stanford Law Review*, 58, 3, 791-846.

Donohue, J. and J. Wolfers, (2006), "The Death Penalty: No Evidence for Deterrence," *Economists' Voice*, April, 1-6.

Draper, D., (1995), "Assessment and Propagation of Model Uncertainty," *Journal of the Royal Statistical Society, series B*, 57, 45-70.

Durlauf, S., (2006), "Assessing Racial Profiling," *Economic Journal*, 116, 515, F402-F426.

Ehrlich, I. and J. Gibbons, (1977), "On the Measurement of the Deterrent Effect of Capital Punishment and the Theory of Deterrence," *Journal of Legal Studies*, 6, 1, 35-50.

Ehrlich, I. and Z. Liu, (1999), "Sensitivity Analyses of the Deterrence Hypothesis: Let's Keep the Econ in Econometrics," *Journal of Law and Economics*, XLII, 455-87.

Ehrlich, I., (1975), "The Deterrent Effect of Capital Punishment: A Question of Life and Death," *American Economic Review*, 65, 397-417.

Ehrlich, I., (1977), "Capital Punishment and Deterrence: Some Further Thoughts and Additional Evidence," *Journal of Political Economy*, 85, 741-88.

Fagan, J., (2006), "Death and Deterrence Redux: Science, Law and Causal Reasoning on Capital Punishment," *Ohio State Journal of Criminal Law*, 4, 1, 255-320.

Fernandez, C., E. Ley, and M. Steel, (2001), "Model Uncertainty in Cross-Country Growth Regressions," *Journal of Applied Econometrics*, 16, 563-576.

Gaus, G., (2006), "Social Complexity and Evolved Moral Principles," mimeo, University of Arizona.

Hansen, B., (2006), "Least Squares Model Averaging," *Econometrica*, forthcoming.

Hansen, L. and T. Sargent, (2006), *Robustness*, manuscript, forthcoming Princeton: Princeton University Press.

Hjort, N. and G. Claesskens, (2003), "Frequentist Model Averaging Estimators," *Journal of the American Statistical Association*, 98, 464, 879-899.

Hoeting, J., M. Clyde, D. Madigan and A. Raftery, (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382-401.

Hjalmarsson, R., (2006), "Does Capital Punishment Have a "Local" Deterrent Effect on Homicides?," mimeo, University of Maryland.

Katz, L., S. Levitt, and E. Shustorovich, (2001), "Prison Conditions, Capital Punishment, and Deterrence," *American Law and Economics Review*, 5, 318-43.

Klein, L., B. Forst, and V. Filatov, (1978), "The Deterrent Effect of Capital Punishment: An Assessment of the Evidence," in *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*, Washington, D.C.: National Academy of Sciences.

Leamer, E., (1978), *Specification Searches*. New York: John Wiley and Sons.

Leamer, E. (1983), "Let's Take the Con out of Econometrics," *American Economic Review*, 73, 31-43.

Liebman, J. S., J. Fagan, & V. West. 2000. Capital attrition: Error rates in capital cases, 1973-1995. *Texas Law Review*, 78, 1839-1861.

McAleer, M. and M. Veall, (1989), "How Fragile Are Fragile Inferences? A Re-Evaluation of the Deterrent of Capital Punishment," *Review of Economics and Statistics*, 71, 99-106.

McManus, W. (1985), "Estimates of the Deterrent Effect of Capital Punishment: The Importance of the Researcher's Prior Beliefs," *Journal of Political Economy*, 93, 2, 417-25.

Mocan, N. and R. Gittings, (2001), "Getting Off Death Row: Commuted Sentences and the Deterrent Effect of Capital Punishment," *Journal of Law and Economics*, 46, 2, 453-78.

Passell, P. and J. Taylor. (1977), "The Deterrent Effect of Capital Punishment: Another View," *American Economic Review*, 85, 445-58.

Pesaran, M. H. and R. Smith, (1994), "A Generalized  $R^2$  Criterion for Regression Models Estimated by the Instrumental Variables Method," *Econometrica*, 62, 3, 705-710.

Raftery, A., (1995) "Bayesian Model Selection in Social Research (with discussion)," in *Sociological Methodology 1995*, P. Marsden ed., (Cambridge, MA: Blackwell).

Raftery, A., D. Madigan, and J. Hoeting, (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association* 92, 437, 179-91.

Rubin, P., (2006), "Reply to Donohue and Wolfers on the Death Penalty and Deterrence," *Economists' Voice*, April.

Sala-i-Martin, X., G. Doppelhofer, and R. Miller, (2004), "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," *American Economic Review*, 94, 4, 813-835.

Shepherd, J., (2005), "Deterrence versus Brutalization: Capital Punishment's Differing Impact Among States," *Michigan Law Review* 104: 203.

Sunstein, C. and A. Vermeule, (2005), "Is Capital Punishment Morally Required? Acts, Omissions, and Life-Life Tradeoffs," *Stanford Law Review*, 58, 3, 703-750.