Model Uncertainty and the Deterrent Effect of Capital Punishment

Ethan Cohen-Cole, *Federal Reserve Bank of Boston*, Steven Durlauf, *University of Wisconsin*, Jeffrey Fagan, *Columbia University School of Law*, and Daniel Nagin, *Carnegie Mellon University*

The reintroduction of capital punishment in 1976 that ended the four-year moratorium on executions generated by the Supreme Court in the 1972 decision *Furman v. Georgia* 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 that 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, weak. (*JEL* G31, G34, D82, C5, K4)

The Department of Justice's National Institute of Justice has provided financial assistance. We are grateful for research assistance provided by Jonathan Larson. Durlauf thanks the National Science Foundation for financial support. Two anonymous referees and the editor John Pepper have provided valuable suggestions. 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.

Send correspondence to: Steven Durlauf, Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706–1393, USA; E-mail: sdurlauf@ssc.wisc.edu.

American Law and Economics Review
doi:10.1093/aler/ahn001
Advance Access publication April 15, 2008

1. Introduction

This paper explores the strength of evidence on the deterrent effect of capital punishment with specific attention to recent studies on 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 systematically address model uncertainty; we offer suggestions on how model uncertainty can be addressed. In terms of substance, we perform a set of exercises 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, after a four-year moratorium on death sentences and executions resulting from the US Supreme Court Decision in *Furman v. Georgia*.[1] Not all states reintroduced capital punishment at the same time when the decision in *Gregg v. Georgia*[2] lifted the Constitutional barrier: they acted at different times to restore capital punishment and have used it at widely varying rates ever since. 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 deterrent effects is fragile, in that it 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 deterrent effects are heterogeneous, so that important properties are masked by imposing a single measure on the statistical

1. 408 U.S. 283 (1972).
2. 428 U.S. 153 (1976).

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 deterrent effect when the number of executions exceeds some empirically identified threshold. Hjalmarsson (2006) explores whether executions have short-run local deterrent 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 previous 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 and update 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. These new studies are also similar in that they all employ aggregate observational data on murder, punishment risks, and rich sets of covariates to evaluate deterrent effects. Despite these communalities, estimates of deterrent 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 a conceptually 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 percent or 0 percent.

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 deterrent 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. The first step in the procedure is the formulation of a space of candidate models. This step involves judgment; there is no algorithm for determining what models should be considered. Second, each model in the model space is estimated. Third, these estimated models are averaged, where the averaging weights represent the probabilities that each of the models is, in fact, the true one. This procedure in essence makes statistical claims that are conditional on a model space rather than on a particular model. One may take either a frequentist or a Bayesian approach to model averaging as its key features involve accounting for model uncertainty and are not predicated on a particular statistical philosophy.

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) 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 [capital punishment] provide no useful evidence on the deterrent evidence of capital punishment" (Blumstein, Cohen, and Nagin, 1978, p. 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.[3] 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.[4] 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) (denoted DRS) and Donohue and Wolfers (2005) (denoted DW).

3. And of course, for large model spaces, chance alone may virtually ensure that there is at least one pair of models with opposite implications, so that regardless of the context, no policy change is made.

4. See McAleer and Veall (1989) for an earlier criticism of Leamer's use of extreme bounds analysis in evaluating deterrence evidence.

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 weak. This will be true in two complementary senses. First, model uncertainty renders calculations of net lives saved from executions very imprecise. While our point estimate of net lives saved from an additional execution is positive, we find that for approximately one-fourth of the models we study, the deterrent effect is negative, i.e., more executions are associated with more homicides. Second, the signs of the relevant deterrence variables are in some cases inconsistent with the deterrence argument, thereby undermining its logic and leading to the possibility that the equilibrium effects of a capital punishment regime can raise the number of homicides.

It is important to recognize that our analysis is not designed to address criticisms of deterrence studies that focus on simply challenging a particular specification. For example, we do not question the validity of the instruments that are employed by DRS even though their validity has been subjected to critique by DW. The reason for this decision is that we wish to understand how model specification choices determine the different findings in the papers. In this sense, we take the behavioral assumptions implicit in the original DRS analysis seriously, and ask to what extent their conclusions depend on details of the statistical instantiation of these assumptions. Arguments about instrumental variable validity typically function outside the logic of a given model.[5] As such, they rely on judgments that are outside the scope of our study.

This focus also means that we restrict our consideration of model uncertainty to the minimum amount needed to encompass DW and DRS. For example, we accept the DW decision to focus on differences between California, Texas, and the rest of the United States rather than pursue a systematic examination of heterogeneity across all the states.[6] Similarly, where DRS and DW coincide, as they do on most issues of measurement, we

5. In other work with Salvador Navarro and David Rivers, one of us (Durlauf) has found that the DRS instruments pass the Stock-Yogo overidentification test. So at least from this dimension, the DRS instruments are vindicated. Of course, this does not address arguments of the type made by DW that the instruments are a priori invalid. DW are in fact clear that they do not provide a formal test of validity. Rubin's (2006) reply to their criticism does not address its substance.

6. DW (p. 826) emphasize that their analyses, which exclude California and Texas, are done in order to evaluate robustness. In our view, the exclusion of Texas is of particular interest because of the disproportionate number of executions there; see Berk (2005) and

employ their common assumptions even though these may be problematic. For example, Cook and Ludwig (2006) argue that the suicide rate with guns is a better measure of gun access than National Rifle Association (NRA) membership. Other examples include heterogeneity in the crimes measured as murder and the specification of error structure in the panel under study.[7] Nor do we address the microfoundations of the DRS regression; Durlauf, Navarro, and Rivers (2007) show that there are interpretation problems involved in moving from individual crime choice problems to the type of regression employed in DRS. While all of these problems are of interest, and at least some may be addressed using model averaging methods,[8] they are outside the DRS/DW comparison. A more radical expansion of the model space would not allow us to understand the differences between the two papers.

In light of the plethora of criticisms of capital punishment deterrence regressions, why do we focus on the disparate results of DRS and DW under alternative statistical specifications? We have two reasons. First, we feel it is of interest to demonstrate how one can move beyond differences in statistical assumptions and corresponding differences in inferences to uncover the information content of a data set given a model space. Second, we feel that model selection exercises fail to provide appropriate information for policy evaluation, when they are used to determine a single model for analysis; more on this follows. Third, we think it is important to understand how empirical findings depend on modeling assumptions of various types. In our view, it is one thing to find that the evidence for a capital punishment deterrent effect depends on whether one believes an instrument is valid and quite another to find it depends on what is assumed about parameter heterogeneity. The former is amenable to social science reasoning in a way that the latter is not. And it seems reasonable that a policymaker's assessment of the desirability of a policy can depend on the source of uncertainty about its impact.

our discussion below. It is not obvious why California should be singled out. We note that theory does not require that the parameters for aggregate deterrence regressions are identical across states.

7. We thank the anonymous referees for raising these issues and for the reference.

8. While we are studying how model averaging can be used to address measurement problems, we are not sure about how this would be done. We mention this as we wish to make clear that model averaging is a tool, not a panacea.

Given the politically charged nature of any claim concerning capital punishment and deterrence, it is important to be explicit about 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 effect made on the basis of these 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 those which have 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 punishment in 1976 following the U.S. Supreme Court's decision in *Gregg v. Georgia*.[9] The empirical question under study is whether a particular policy regime, one in which the exercise of capital punishment is rare, has produced deterrence, nothing more.[10] We also do not deal with broader issues of whether regressions of the type we study can be interpreted as providing causal inferences.

The following section outlines our methodology. The section "Dezhbakhsh, Rubin, and Shepherd versus Donohue and Wolfers" describes the studies by Dezhbakhsh, Rubin, and Shepherd and Donohue and Wolfers, which will form the basis of our empirical analysis. The section that follows discusses implementation issues. The next section reports results, which is followed by the section presenting the summary and conclusions.

2. Model Uncertainty and Model Averaging: General Principles

We follow the discussion in Brock, Durlauf, and West (2003) to describe our statistical framework. Let δ denote the measure of the deterrent effect

9. For this reason, the assertion by Mocan and Gittings (2006) that DW's claim that the current death penalty regime has no deterrent effect is at odds with the basic logic of the choice-based model of crime is fallacious.

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

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 δ given a single model, and so one finds reports of estimates of δ 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 δ 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 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.[11] One of the sources of the different findings in DRS and DW is the choice of data to use. DW argue that excluding a single state, such as California or Texas, from the DRS 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 Texas and one that does not. How do we propose adjudicating the disagreement? We argue that one should construct a weighted average of the results from each study, where the weights are

11. These types of disagreements can be formalized using the probabilistic notion of exchangeability; see Brock and Durlauf (2001) for discussion.

model probabilities.[12] DRS can be interpreted as placing a prior probability of 1 on the model with data that includes California and Texas whereas a researcher using DW would place a prior probability of 1 on the model without California and Texas. (To be clear, DW themselves do not endorse any particular model; rather they use it to illustrate the fragility of the claims in DRS.) 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, δ, which measures the effects of a policy. In the context of the capital punishment literature, δ 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 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: $\mu(m \mid D)$. These posterior probabilities allow us to average the model-specific estimates to produce an estimate that accounts for the model uncertainty:

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

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

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

The estimate $\hat{\delta}_M$ thus accounts for the information contained in each specific model about δ and weights this information according to the likelihood that 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 policy questions such as whether to implement capital punishment in a

12. This can be thought of simply as the conditional probability that a given model describes the data. We discuss this at greater length in the following.

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 δ.

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 δ 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 any 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 $\text{var}(\hat{\delta}_m) = 0\ \forall m$, 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 δ.

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

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

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 is a useful benchmark, 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 δ. 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 its superiority with respect to some goodness of fit criterion. 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 disaggregated data set to have been used to study deterrence and compares favorably with other data sets that have subsequently been used.

The DRS 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 ten to nineteen-year-olds and twenty to twenty-nine-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 DRS murder rate regression

is

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

The overall deterrent effect of capital punishment can be evaluated by viewing the parameters δ_1, δ_2, and δ_3. Estimates from DRS and DW use δ_3 to determine the number of lives saved or lost from executions; we follow this method.[13] Since each of the variables associated with these parameters is endogenous, these parameters are estimated using instrumental variables. The three deterrence variables are assumed to follow

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

$$\frac{DeathSentences_{s,t}}{HomicideArrests_{s,t}} = \theta_0 + \theta_1 \frac{Murders_{c,s,t}}{pop_{c,s,t}} + \theta_2 JudicialExpense_{s,t} + \sum_{j=1}^{6} \theta_{3,j} PartisanInfluence_{s,j,t} + \theta_4 Admissions_{s,t} + \sum_t \theta_{5,t} Time_t + \varepsilon''_{c,s,t}, \quad (6)$$

13. "Net lives saved" is calculated as the number of lives saved as the result of one execution; formally, $NetLivesSaved = \delta_3(population/100,000) \times (1/\#Executions)$, where δ_3 is the coefficient on the execution variable. Including (scaled) population allows conversion from the per capita murder rate. Population numbers are from 1996.

and

$$\frac{Executions_{s,t}}{DeathSentences_{s,t}} = \theta_0 + \theta_1 \frac{Murders_{c,s,t}}{pop_{c,s,t}} + \theta_2 JudicialExpense_{s,t} + \sum_{j=1}^{6} \theta_{3,j} PartisanInfluence_{s,j,t} + \sum_{t} \theta_{4,t} Time_t + \varepsilon'''_{c,s,t}. \quad (7)$$

Instruments resolve the simultaneity between (4) and (5)–(7). We follow DRS and DW in estimating Equations (5)–(7) and then using predicted values in (4) to derive estimates of the relevant deterrence coefficients. Instruments include police payroll, judicial expenditures, six partisan influence variables, and prison admissions. Though one might have some concern over the use of these instruments, we take their appropriateness as given while turning our attention to specification. The variable $Pop_{c,s,t}$ indicates the population in county c, state s, and time t, divided by 100,000. $PartisanInfluence_{s,j,t}$ is measured by the Republican presidential candidate's vote share in the most recent election. The Republican vote share in the most recent election is multiplied by a dummy indicator for the election. Thus these variables appear individually, and associated coefficients are indexed by election j. Finally, *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).[14] Additional information is available in the original study.

DRS (pp. 362–63) present the results from this system of equations given six different versions of the variable $\frac{Executions}{DeathSentences}$. They consider three different measures of execution probabilities in the six columns of their tables 3 and 4; these are

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

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

14. DRS 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.

Table 1. Dezhbakhsh, Rubin, and Shepherd / Donohue and Wolfers Differences

	DRS Baseline	DW Version 1	DW Version 2	DW Version 3
PartisanInfluence	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 data set. 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.

and

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

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 a death sentence occurred. Table 7 (p. 824) in DW 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 DW 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 DRS findings on the grounds of fragility. Specifically, they show that certain modifications to the DRS statistical model can strongly affect findings of a strong deterrent effect to capital punishment. First, they constrain the partisan influence variables in (5)–(7) so that rather than use six distinct regressors, which allow for each election to have a distinct effect on crime, they simply use the sum of the DRS variables as a single regressor. Second, they omit California and Texas from the analysis. These changes cause the sign of the estimated deterrent effect to reverse. That is, under these alternative specifications, each execution is predicted to *increase* the number of murders. Table 1 summarizes the different assumptions in the two papers.

The differences between the DRS and DW findings illustrate how model uncertainty matters for substantive empirical claims, even when this uncertainty is predicated on a common social science theoretical structure.

Each of these papers takes a particular stand on the instrumental variables to be employed and the interstate 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

4.1. 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. DW (2005), who successfully replicate the results in DRS, 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 for assistance when we were unable to replicate their results using data they had earlier supplied to one of us.

The data are a panel of state- and county-level data covering the time period 1977–1996. The FBI was the source of information on crime and arrest rates. The BJS was the source of data 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 were 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. Following DRS, we 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 and are from the Bureau of Economic Analysis. NRA membership rates come directly from the National Rifle Association.

4.2. Model Space Construction

Conceptually, our approach to reconciling different empirical studies is to regard the study-specific models as elements of a larger model space and to ask what inferences one can draw when one does not wish to assume that a particular element of the space is the correct specification. How should one think about the model space? Our approach is to use the assumptions that differentiate DRS and DW as the basis for model space construction. In essence, we consider the different assumptions of the two papers and consider all possible combinations of the assumptions that are internally consistent. For example, DRS assume that data from California and Texas are exchangeable with the rest of the country, whereas DW assume this is not the case. One can augment the initial two models with one where California is assumed exchangeable and Texas is not, and vice versa. Thus model uncertainty with respect to exchangeability produces four possible sets of assumptions. Similarly, one can identify different IV assumptions that are generated by the differences between DRS and DW. When different exchangeability and IV assumptions are combined, this produces the model space. Notice that this approach is conservative in that it does not relax assumptions that are common to both DRS and DW.

To implement model averaging using available methods, specifically those developed in Raftery, Madigan, and Hoeting (1997), we translate the differences between DRS and DW into differences in the choice of regressors at different stages of the analysis. With respect to the handling of the partisanship instrumental variables, we do this by creating a set of candidate variables: five of the DRS variables (Republican vote in previous election multiplied by a dummy for the previous election) and the DW partisanship variable (sum of the DRS vote variables).[15] This choice of candidate instruments spans both DW and DRS, since the DRS specification is contained in the linear span of our six candidate variables. Thus DRS and DW may be understood as making different instrumental variable choices. 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 DW. 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

15. The DRS variable corresponding to the 2000 election is omitted.

California, and corresponding variables that are the products of the deterrent variables with a dummy for Texas. Note that this approach to dealing with California and Texas is substantively different from DW in that the California and Texas data are retained and will affect all model parameter coefficients. An advantage of our approach to modeling the DRS and DW differences for instrumental variables and Texas/California comparability is that the differences between the two papers can be interpreted as reflecting different assumptions about parameter heterogeneity and that, further, the heterogeneity is modeled by including additional regressors in a baseline regression. Notice that from this perspective, DRS assume greater heterogeneity than DW in the deterrence variable equations, whereas DW assume greater heterogeneity in the crime equation.

Each candidate model is therefore defined as a choice of variables for the system of equations (4)–(7). Different instrumental variables generate different specifications of (5)–(7), and different choices with respect to heterogeneity between California, Texas, and the rest of the United States generate different specifications of (4). Our model space consists of 12,288 different specifications. This figure may be decomposed as follows. In Equation (4), we include all combinations of California and Texas dummies interacted with the three deterrence variables. This produces six variables and sixty-four possible models (2^6 combinations).[16] Within the instrument space, we have six vote variables, as specified above, as well as variables for police payroll, judicial expenses, and prison admissions. In order to achieve identification, we restrict the space such that at least four instruments are included, three of which always include the policy, judicial, and prison admit variables. We also always include the DW variable that encompasses all six of the DRS variables. This produces $2^5 = 32$ specifications. By including the DW variable in all the specifications, we parallel the way in which

16. 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 two *sets* of variables enter the model space. The two variable sets are those that correspond to the three deterrence variables interacted with one of the state dummies. Thus, in this version, the model space effectively includes two new "elements," each of which consists of three variables. As the analyses with this 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 brevity.

uncertainty about heterogeneity with respect to California and Texas is modeled as we treat the model uncertainty as surrounding parameter heterogeneity and treat the heterogeneity as a question of regressor inclusion. Thus, for each of the six dependent variable choices from DRS, we have $32 \times 64 = 2048$ models. Our grand average thus contains the full 12,288. The reduced IV and exchangeability cases contain 32 and 64 models, respectively.

4.3. Model Weight Calculation

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

$$p(m|D) \approx \exp\left(-\frac{1}{2}BIC_k\right) \Big/ \sum_{l=1}^{K} \exp\left(-\frac{1}{2}BIC_l\right) \tag{11}$$

where *BIC* is defined as

$$BIC = n \log\left(1 - R_p^2\right) + p \log n \tag{12}$$

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

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 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

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.

Dependent Variable: Annual Homicides per 100,000 Residents						
	(1)	(2)	(3)	(4)	(5)	(6)
Probability of arrest (δ_1)	1.26	1.29	1.27	1.65	−2.68	2.02
	(0.24)	(0.21)	(0.24)	(8.78)	(6.19)	(8.70)
Probability of death sentence given arrest (δ_2)	−25.05	−15.14	−36.61	96.81	78.07	137.64
	(20.12)	(13.84)	(21.95)	(100.15)	(25.13)	(120.68)
Probability of execution given death sentence (δ_3)	0.27	−3.82	1.62	−15.60	−3.30	−18.96
	(6.12)	(4.16)	(4.49)	(11.36)	(3.12)	(12.78)
Net lives saved	−0.96	28.36	−10.60	112.75	24.69	136.85
	(41.86)	(17.02)	(27.17)	(177.45)	(47.03)	(201.05)

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 ten to nineteen-years old, twenty to twenty-nine years old; black, white, or other; male or female; state NRA membership. Two-stage least squares estimation is employed. Net lives saved evaluated for a state with the characteristics of the average death penalty state in 1996. Instruments include state-level police payroll, judicial expenditures, Republican vote shares, and prison admissions.

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 one-half the value of the associated parameter estimate, large otherwise. This roughly corresponds to the treatment of *t*-statistics greater than two as statistically significant.

5. Results

5.1. Averaging

Recall that DRS report deterrence findings for six different constructions [equations (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 four of the six categories

(Columns 2, 4, 5, and 6), the estimated number of net lives saved is positive, with a standard error of approximately the same size or larger than the estimate. For the other two categories the net lives saved estimate is negative, but with large associated standard errors. This inconsistency contrasts with the DRS 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 DW in the sense that for some categories, they found that alternative assumptions led to a statistically significant association of additional murders for each execution. But this does not contradict the claim of DW that DRS's findings are model-specific. Our averaging exercise reinforces the DW claim. It demonstrates that their conclusion is not an artifact of their having reported a particularly unfavorable alternative specification relative to the DRS baseline.

Some additional evidence of the lack of firm deterrence evidence may be obtained from a consideration of the three deterrence parameters δ_1 (arrest probability), δ_2 (death sentence probability given arrest), and δ_3 (probability of execution given death sentence). In principle, the choice-theoretic logic underlying the murder rate regressions implies that each of the parameters is negative. Our model-averaged estimates 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 columns 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 net lives, the deterrence coefficients δ_1 and δ_2 do not have the expected signs (though each has a large standard error). In fact, for none of the six columns do we find a case where all three deterrence variables have negative signs.

This absence of a pattern of deterrence coefficient estimates with negative signs not only illustrates the weakness of the evidence embodied in the data, but also indicates that the effect of capital punishment on net lives saved may differ from the conventional calculation as described in footnote 7. The net lives saved calculations are made based on a transformation of the coefficient on the probability of an execution given the death sentence. However, the equilibrium effect of capital punishment, as a policy regime,

Table 3. Deterrent Effects: Full Averaging

Dependent Variable: Annual Homicides per 100,000 Residents	
Probability of Arrest (δ_1)	0.80
	(4.06)
Probability of Death Sentence Given Arrest (δ_2)	39.28
	(50.31)
Probability of Execution Given Death Sentence (δ_3)	−6.63
	(7.00)
Net Lives Saved	48.51
	(38.72)

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 ten to nineteen years old, twenty to twenty-nine years old; black, white, or other; male or female; state NRA membership. Two-stage least squares estimation is employed. Standard errors are in parentheses. Net lives saved evaluated for a state with the characteristics of the average death penalty state in 1996. Instruments include state-level police payroll, judicial expenditures, Republican vote shares, and prison admissions. Full averaging is the unweighted grand mean taken over the six columns in table 2.

also depends on the effects on homicide of the probabilities of arrest and the probabilities of a death sentence. Positive coefficients imply that some aspects of the capital punishment regime raise the homicide rate, so one cannot make claims about the overall effect of the regime. While we do not have a clear theory as to why the coefficients are positive,[17] the absence of a causal explanation does little to mitigate the failure of the data to provide strong evidence that an overall deterrence effect is present.

How does the information in the various categories combine to produce an overall deterrence estimate? Table 3 integrates the information across the DRS 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 constructing a new model space and then computing posterior

17. One reason for the anomalous sign may be the high reversal rate on death sentences (see Liebman, Fagan, and West, 2000) (68 percent reversal rate on death sentences). Only about one death sentence in nine survives to the execution stage. If those experiencing reversals would have received more severe punishment in absence of capital punishment, then the finding can be understood. Similarly, if jury behavior is affected by the presence of capital punishment, an issue first raised by Andreoni (1995), anomalous results (from the perspective of the deterrence theory) can be produced. To be clear, these explanations are speculative—our findings are very much a black box.

model probabilities; this allows the easiest comparisons with table 2.[18] As the table indicates, the summary deterrence statistic—expected number of lives saved per execution—is large relative to most studies, forty-nine, but the standard deviation of the estimate is quite high as well, thirty-nine. This suggests that the evidentiary support for a deterrent 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 deterrent effects, with large standard errors for each estimate and a positive sign for the overall estimate of δ_2 (death sentence likelihood given arrest), which, as emphasized above, is in contradiction to what theory predicts. Again, there does not exist a coherent pattern to the underlying deterrence coefficients that is consistent with the theoretical basis for the capital punishment effect.

As we have discussed, the differences between DRS and DW are generated by differences in instrumental variable choices and state heterogeneity assumptions. In order to understand the roles of these two classes of assumptions, we repeat our averaging exercises focusing on two model subsets: one in which all states are used so that the only model uncertainty is generated by instrumental variable choice, and a second where the original DRS instrumental variable structure is retained in all models so that the only model uncertainty concerns the comparability of California and Texas with the rest of the country. These results are reported in table 4 and are most usefully compared with table 2, the model averaging exercises in which the two types of model uncertainty interact to produce the overall model space. As this comparison indicates, the weakness of evidentiary support for deterrence is preserved across both the smaller model spaces. For both of our model space subset exercises, the net lives saved estimate is negative for columns 1 and 3, just as it is for the corresponding columns in table 2.

Table 4 indicates that both sources of model uncertainty—IV choice and state heterogeneity—matter for deterrence claims. The major difference between table 2 and table 4 is found in column 5 of panel B, where the estimated net lives saved is both positive and highly statistically significant. What this means is that even if one fixes the DRS instruments,

18. 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.

Table 4. Analysis of Model Space Subsets

Dependent Variable: Annual Homicides per 100,000 Residents						
	(1)	(2)	(3)	(4)	(5)	(6)
(A) Instrumental Variable Uncertainty						
Probability of arrest (δ_1)	1.32	1.31	1.33	4.48	0.65	4.93
	(0.25)	(0.21)	(0.25)	(6.95)	(4.48)	(7.03)
Probability of death sentence given arrest (δ_2)	−33.25	−16.68	−41.59	105.99	92.92	148.41
	(21.77)	(17.13)	(21.57)	(96.23)	(24.52)	(115.26)
Probability of execution given death sentence (δ_3)	4.38	−0.78	4.53	−14.93	−3.75	−18.43
	(6.82)	(5.28)	(4.74)	(10.99)	(2.89)	(12.46)
Net lives saved	−30.36	6.60	−31.46	107.99	27.90	133.09
	(66.51)	(68.10)	(35.47)	(156.54)	(40.40)	(177.61)
(B) State Heterogeneity Uncertainty						
Probability of arrest (δ_1)	1.11	1.10	1.12	−4.84	−6.56	−4.09
	(0.12)	(0.07)	(0.11)	(2.66)	(2.26)	(2.69)
Probability of death sentence given arrest (δ_2)	−24.90	−12.55	−23.83	17.51	60.09	43.86
	(12.78)	(8.64)	(11.31)	(14.77)	(10.55)	(17.69)
Probability of execution given death sentence (δ_3)	1.47	−1.63	0.99	−10.61	−6.69	−11.98
	(3.85)	(2.45)	(2.58)	(1.70)	(0.60)	(1.85)
Net lives saved	−11.53	10.67	−8.09	75.02	46.93	84.84
	(27.03)	(17.20)	(18.11)	(11.93)	(4.21)	(12.99)

Panel A: 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 ten to nineteen years old, twenty to twenty-nine years old; black, white, or other; male or female; state NRA membership. Two-stage least squares estimation is employed. Standard errors are in parentheses. Net lives saved evaluated for a state with the characteristics of the average death penalty state in 1996. Endogenous independent variables are shown in panel A. Instruments include state-level police payroll, judicial expenditures, Republican vote shares, and prison admissions. To ensure identification, we require a minimum of three instruments in each specification.
Panel B: Controls in homicide equation 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 ten to nineteen years old, twenty to twenty-nine years old; black, white, or other; male or female; state NRA membership. Two-stage least squares estimation is employed. Standard errors are in parentheses. Net lives saved evaluated for a state with the characteristics of the average death penalty state in 1996. Instruments include state-level police payroll, judicial expenditures, Republican vote shares, and prison admissions. Model space is over the combination of the three dummy deterrence variables in the first column above and the state dummies for TX and CA.

for this measure of $\frac{Executions}{DeathSentences}$, model uncertainty associated with the exchangeability of California and Texas with the rest of the country does not vitiate the strength of the DRS deterrence evidence.

Overall, the analysis of this section indicates that evidence of a capital punishment deterrent effect is generally weak. As indicated in table 3, the model-averaged estimates of net lives saved are very imprecise, with a standard error nearly equal to the point estimate. When one allows for different constructions of $\frac{Executions}{DeathSentences}$, as done in table 2, one even finds point estimates suggesting that an additional execution increases the homicide rate, and only one of six estimates is much larger than the associated standard error. Further, the model-averaged estimates of the deterrence parameters are also very imprecisely estimated, and in many cases point estimates are the opposite of what is theoretically implied if the deterrent effect is present.

That said, it is important to emphasize that our point estimates of net lives saved are often large; for example, forty-nine in table 3, as opposed to eighteen, the number usually associated with DRS. In principle, one could argue that if a policymaker is risk neutral, the imprecision of the estimate is irrelevant. But this is not a tenable position. If taken seriously, it implies that any unbiased estimate of a deterrent effect should have the same policy implication as any other. More substantively, we see no reason why a policymaker should be risk neutral. For contexts such as FDA evaluation of a new drug, we cannot imagine a serious claim that the uncertainty associated with estimates of side effects is irrelevant to a decision in favor of approval. Similar considerations apply to capital punishment, and given the complex moral issues associated with it we, believe it natural for a policymaker to trade off the strength of the deterrent effect against other desiderata. And the imprecision of our estimates implies that there is some probability that executions increase the murder rate, which is perfectly compatible with rational behavior on the part of criminals, e.g., witness elimination.

5.2. Properties of the Density of Deterrent Effects

In this section, we consider how deterrence estimates vary across the model space. Our discussion so far has focused on identifying "bottom" line estimates as opposed to understanding the properties of particular model estimates. The density (across the model space) of deterrent effects is of additional interest for several reasons.

The first reason why particular model estimates are interesting is that they can provide some insight into the relationship among DRS, DW, and the model space that is defined by their differing assumptions. In table 5 we compare DRS and DW estimates with the elements of the model space

Table 5. Comparison of Deterrent Effects Estimates

Dependent Variable: Annual Homicides per 100,000 Residents					
	1 Smallest	2 Largest	3 Largest Posterior	4 "best" DRS	5 Average DW
Probability of arrest (δ_1)	1.68	19.96	−7.66	−3.33	−4.29
	(0.02)	(0.58)	(0.68)	(0.52)	(0.54)
Probability of death sentence given arrest (δ_2)	−67.85	393.94	55.62	−32.12	−9.03
	(7.71)	(8.88)	(6.59)	(16.22)	(13.40)
Probability of execution given death sentence (δ_3)	14.84	−44.58	−6.32	−7.40	−0.57
	(1.29)	(1.27)	(0.48)	(0.72)	(0.72)
Net lives saved	−121.2	364.02	51.6	52.0	0.08
	(9.06)	(8.91)	(3.37)	(5.1)	(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 ten to nineteen years old, twenty to twenty-nine years old; black, white, or other; male or female; state NRA membership. Two-stage least squares estimation is employed. Standard errors are in parentheses. Net lives saved evaluated for a state with the characteristics of the average death penalty state in 1996. Instruments include state-level police payroll, judicial expenditures, Republican vote shares, and prison admissions. 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.

that provide the largest and smallest estimates of the deterrent effects. To be precise, Column 1 of table 5 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 DRS specification with the largest net lives saved per execution. This corresponds to column 3 of DRS (pp. 362–63, tables 3 and 4). Column 5 provides a simple average of the DW findings. Recall from table 1 that DW provide three variations on the six specifications of DRS. We average all eighteen of these for the result in Column 5. (We average the DW results since their focus was on the fragility of DRS, not promoting a particular alternative model.)

These results help place the differing conclusions of DRS and DW in context. They suggest that DRS'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. The most favorable model among the limited set they considered is close, in terms of net lives saved, to the model with highest posterior weight of being true.[19] Instead, the message of this comparison is that DRS'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. These additional models contain additional information about deterrence. We would also note that a comparison of column 1 with column 5 illustrates that the DW analysis did not focus on outlier models in the sense that column 1 illustrates a far larger increase in murders than the DW average; this can also be seen in figure 2.

A second reason why the density of deterrent effects across the model space is important is that it provides relevant information when a policymaker is risk averse; standard errors generally do not fully reflect the relevant uncertainty associated with deterrence estimates when preferences are not quadratic. 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 six DRS categories. The weights are the posterior model probabilities. Figure 2 provides the corresponding 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. In the unweighted histogram each model receives identical weight, of course. The figures indicate the locations in the histograms of the DRS and DW model estimates associated with the largest and smallest (negative) number of net lives saved, respectively across those they considered.

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 probability, for all the models in figure 1 which produce a deterrent

19. The model with the greatest posterior is based on the construction of $\frac{Executions}{DeathSentences}$ corresponding to table 2, column 5. It includes five of the six interacted deterrent variables; the exception is the California dummy interacted with the probability of death sentence given arrest.

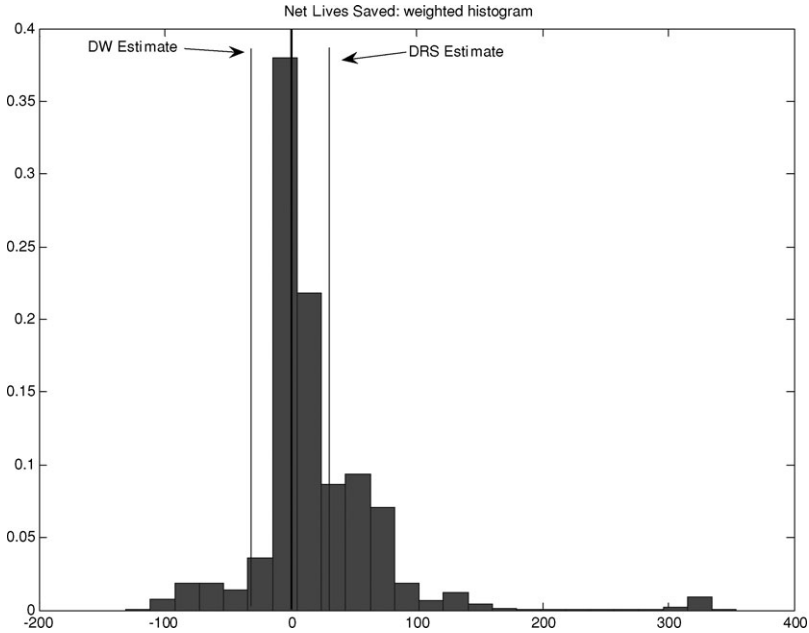

Figure 1. This figure is a weighted histogram of the net lives saved for all the models we have considered, including models for each of the DRS categories. The weights are the posterior model probabilities. The DRS and DW lines correspond to the individual model from each with the largest and smallest number of lives saved, respectively.

effect is 0.72; in figure 2, the corresponding percentage of models with positive deterrent effects is also 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 is a set of models in the unweighted histogram that 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.

Finally, table 6 provides results on cumulated model probabilities and associated deterrent effects; the table thus provides aggregate statistics based on the histogram in figure 1. As the table indicates, the probability that an additional execution increases the murder rate by twenty or more is 15 percent whereas the probability that an additional execution decreases the murder rate by twenty or more is 53 percent. This reflects the asymmetry

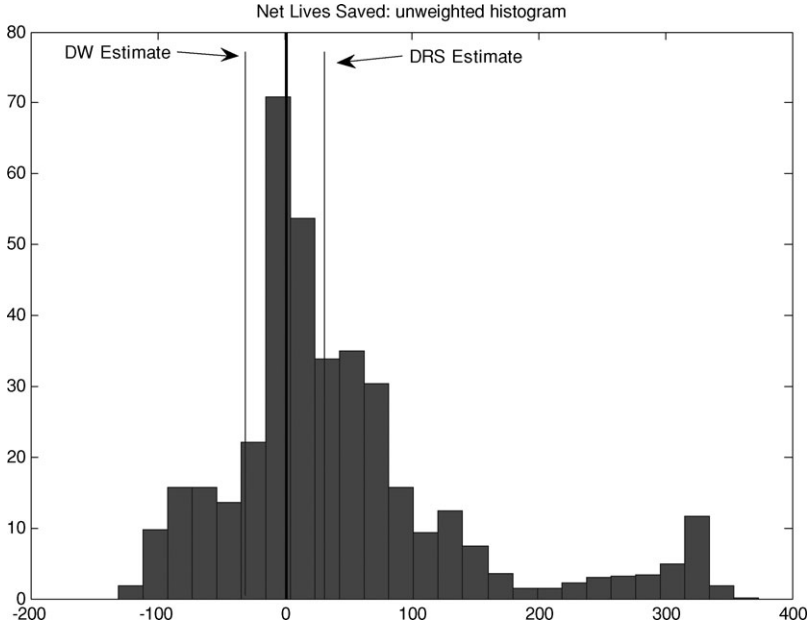

Figure 2. This figure provides the same histogram as figure 1 when the models are all assigned equal weights. The DRS and DW lines correspond to the individual model from each with the largest and smallest number of lives saved, respectively.

Table 6. Cumulated Histogram Probabilities

Sum of Model Probabilities, with net lives saved										
< -20	< -10	< -5	< -2	< -1	<0	<1	<2	<5	<10	<20
0.15	0.18	0.20	0.25	0.26	0.28	0.28	0.29	0.32	0.37	0.47

Model probabilities are calculated as specified in the text. Each model generates a coefficient for probability of execution conditional on a death sentence and an associated net lives saved value. Results for this exercise use full model space as specified in the text. Each column indicates the percentage of summed model probabilities for models with net lives saved <X, with X as indicated in the table. 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 ten to nineteen years old, twenty to twenty nine years old; black, white, or other; male or female; and state NRA membership.

in the histogram in figure 1. Of particular interest are the probabilities for values near zero. The probabilities for models that are associated with net lives saved of five and fewer is approximately one-third. So, while the DRS estimate of eighteen is relatively close to (and indeed smaller than) the

median of the histogram, twenty-four, there is substantial uncertainty about the actual deterrence level.

6. Conclusions

The empirical study of deterrent effects is an example of a problem domain where theory, economic or otherwise, does not provide strong guidance on how to construct a statistical model that maps theory to empirical work. This openendedness 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 DRS and DW leads us to support the conclusion that DRS claims about strong deterrent 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 DW regressions (although 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 thirty-six) but imprecisely estimated (weighted average standard deviation twenty-six). An analysis of the behavior of deterrence estimates across the model space indicates that, while the aggregate probability associated with models that produce a deterrent effect is 0.72, the individual estimates vary widely and include a nontrivial probability that there is a large increase in homicides associated with execution (probability 0.15 of twenty or more homicides). These conclusions may seem frustrating, since they make clear that our conclusion is not that there is no deterrent effect present, but rather that inferences on its magnitude are so imprecise as to make representation of strong claims impossible. 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 in but a handful of states;

in particular, uncertainty about the comparability of Texas to other states 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 0.72. However, 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, by themselves, provide guidance as to how a policymaker should use these numbers when comparing policy choices. Should a policymaker care whether 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 contain a risk aversion component that is relevant when assessing deterrent effects. In fact, 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 specific 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.[20] For our purposes, they suggest that simply computing expected deterrent effects is inadequate.

Further progress in evaluating deterrence, in our view, requires questions of criminal policy evaluation be considered 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 which incorporate a richer set of factors than simply minimizing the number of murders.[21] 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 the 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)

References

Andreoni, J. 1995. "Criminal Deterrence in the Reduced Form: A New Perspective on Ehrlich's Seminal Study," 33(3) *Economic Inquiry* 476–83.

Baldus, D., and J. Cole. 1975. "A Comparison of the Work of Thorsten Sellin and Isaac Ehrlich on the Deterrent Effect of Capital Punishment," 85 *Yale Law Journal* 170–84.

20. 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.

21. Sunstein and Vermeule's analysis focuses on evaluating capital punishment when it leads to a reduction of the murder rate, and does not discuss the appropriate levels of risk or ambiguity aversion on the part of the policymaker. They also largely ignore any competing ethical claims on policy choices.

Berk, R. 2005. "New Claims about Executions and General Deterrence: Deja Vu All Over Again?," 2(2) *Journal of Empirical Legal Studies* 303–30.

Blumstein, A., J. Cohen, and D. Nagin. 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," 85 *Yale Law Journal* 187–208.

Brock, W., and S. Durlauf. 2001. "Growth Economics and Reality," 15 *World Bank Economic Review*, 229–72.

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

Brock, W., S. Durlauf, J. Nason, and G. Rondina. 2007. "Simple Versus Optimal Rules as Guides to Policy," 54(5) *Journal of Monetary Economics* 1372–1396.

Cook, P., and J. Ludwig. 2006. "The Social Costs of Gun Ownership," 90(1–2) *Journal of Public Economics* 379–91.

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

Dezhbakhsh, H., and J. Shepard. 2006. "The Deterrent Effect of Capital Punishment: Evidence from a 'Judicial Experiment'," 44(3) *Economic Inquiry* 512–35.

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

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

Doppelhofer, G., and M. Weeks. 2007. *Jointness of Growth Determinants*. Mimeo, University of Cambridge, UK.

Draper, D. 1995. "Assessment and Propagation of Model Uncertainty," 57 *Journal of the Royal Statistical Society, Series B* 45–70.

Durlauf, S. 2006. "Assessing Racial Profiling," 116(515) *Economic Journal* F402–26.

Durlauf, S., D. Rivers, and S. Navarro. 2007. *Notes on the Econometrics of Deterrence*. Mimeo, University of Wisconsin.

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

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

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

Ehrlich, I., and Z. Liu. 1999. "Sensitivity Analyses of the Deterrence Hypothesis: Let's Keep the Econ in Econometrics," 15(2) *Journal of Law and Economics* 455–87.

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

Fernandez, C., E. Ley, and M. Steel. 2001. "Model Uncertainty in Cross-Country Growth Regressions," 16 *Journal of Applied Econometrics* 563–76.

Gaus, G. Forthcoming. *"Social Complexity and Evolved Moral Principles," in Liberalism, Conservatism, and Hayek's Idea of Spontaneous Order, edited by P. McNamara*, London: Palgrave Macmillan.

Hansen, B. 2007. "Least Squares Model Averaging," 75(4) *Econometrica* 1175–1189.

Hansen, L., and T. Sargent. 2007. *Robustness*. Manuscript. Princeton, NJ: Princeton University Press.

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

Hjort, N., and G Claesskens. 2003. "Frequentist Model Averaging Estimators," 98(464) *Journal of the American Statistical Association* 879–99.

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

Katz, L., S. Levitt, and E. Shustorovich. 2001. "Prison Conditions, Capital Punishment, and Deterrence," 5 *American Law and Economics Review* 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, DC: National Academy of Sciences.

Leamer, E. 1978. *Specification Searches*. New York: John Wiley.

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

Liebman, J. S., J. Fagan, and V. West. 2000. "Capital Attrition: Error Rates in Capital Cases, 1973–1995," 78 *Texas Law Review* 1839–1861.

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

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

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

Mocan, N., and R. Gittings. 2006. "The Impact of Incentives on Human Behavior: Can We Make it Disappear? The Case of the Death Penalty," *National Bureau of Economic Research Working Paper no. 12631*, Cambridge, MA.

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

Pesaran, M. H., and R. Smith. 1994. "A Generalized R^2 Criterion for Regression Models Estimated by the Instrumental Variables Method," 62(3) *Econometrica* 705–10.

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

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

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

Sala-i-Martin, X., G. Doppelhofer, and R. Miller. 2004. "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," 94(4) *American Economic Review* 813–35.

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

Sunstein, C., and A. Vermeule. 2005. "Is Capital Punishment Morally Required? Acts, Omissions, and Life-Life Tradeoffs," 58(3) *Stanford Law Review* 703–50.